New statistical perspectives on efficient Big Data algorithms for high-dimensional Bayesian regression and model selection

Daniel Christian Ahfock

Fitzwilliam College
MRC Biostatistics Unit
University of Cambridge

September 2018

A thesis presented for the degree of
Doctor of Philosophy

Abstract

This thesis is focused on the development of computationally efficient procedures for regression modelling with datasets containing a large number of observations. Standard algorithms be prohibitively computationally demanding on large n datasets, and we propose and analyse new computational methods for model fitting and selection. We explore three different generic strategies for tall datasets. Divide and conquer approaches split the full dataset into subsets, with the subsets then being analysed independently in parallel. The subset results are then pooled into an overall consensus. Subsampling based methods repeatedly use minibatches of data to estimate quantities of interest. The third strategy is sketching, a probabilistic data compression technique developed in the computer science community. Sketching uses random projection to compress the original large dataset, producing a smaller surrogate dataset that is less computationally demanding to work with. The sketched dataset can be used for approximate inference. We test our regression algorithms on several large n genetic datasets, aiming to find associations between genetic variants and red blood cell traits.

Bayesian divide and conquer and subsampling methods have been studied in the fixed model setting but little attention has been given to model selection. An important task in Bayesian model selection is computation of the integrated likelihood. We propose divide and conquer and subsampling algorithms for estimating the integrated likelihood. The divide and conquer approach is based on data augmentation, which is particularly useful for logistic regression. The subsampling approach involves constructing upper and lower bounds on the integrated likelihood using information theory.

Sketching algorithms generate a compressed set of responses and predictors than can then be used to estimate regression coefficients. Sketching algorithms use random projections to compress the original dataset and this stochastic generation process makes them amenable to statistical analysis. We examine the statistical properties of sketching algorithms, which allows us to quantify the error in the coefficients estimated using the sketched dataset. The proportion of variance explained by the model proves to be an important quantity in choosing between alternative sketching algorithms. This is particularly relevant to genetic studies, where the signal to noise ratio can be low.

We also investigate sketching as a tool for posterior approximation. The sketched dataset can be used to generate an approximate posterior distribution over models. As expected, the quality of the posterior approximation increases with the number of observations in the sketched dataset. The trade-off is that computational cost of sketching increases with the size of the desired sketched dataset. The main conclusion is that impractically large sketch sizes are needed to obtain a tolerable approximation of the posterior distribution over models. We test the performance of sketching for posterior approximation on a large genetic dataset. A key finding is that false positives are a major issue when performing model selection.

Practical regression analysis with large n datasets can require specialised algorithms. Parallel processing, subsampling and random projection are all useful tools for computationally efficient regression modelling.

Declaration

- This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specified in the text.
- I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution.
- Chapters 4 and 5 are joint work with Sylvia Richardson and William Astle. The contents in these chapters have been submitted together jointly for publication.

Acknowledgement

Firstly, thank you to my supervisors Sylvia Richardson and William Astle for all of their help and guidance over the course of the PhD. The work in this thesis would not have come together without their efforts. I am also very grateful to my parents, Tony and Georgette, and to my sister Melody for all of their support and encouragement. Even though Cambridge is a long way from Australia, I never felt too far away from home. Finally, thank you to my wonderful wife Amy for being such an amazing partner on this journey.

Contents

List of Figures

List of Tables

ii

1

Introduction

In their recent monograph 'Computer age statistical inference', Efron and Hastie (2016) trace the development of statistical methodology over the past century with a particular emphasis on computationally intensive procedures. A central theme in the book is that the frontiers of statistical inference are pushed forward as both available datasets and data analysis methods evolve in complexity and richness. The capacity to acquire data and the capacity to analyse data typically advance in tandem, jointly propelled by technological developments. New systems create datasets of unprecedented size and scope that require analysis with new computer driven statistical procedures. For example, microarray technology presented large scale-hypothesis testing challenges. Computationally intensive false discovery rate methodology (Benjamini and Hochberg, 1995) was then honed in the analysis of this data. In turn, these new algorithms open the doors for more ambitious scientific investigations. False discovery rate methods can serve as a valuable tool in genome-wide association studies. New datasets can strain existing methodology, and innovation can spring from these difficulties. The emergence of Big Data marks another period where statistical methodology will simultaneously be challenged and afforded the opportunity to expand its reach.

Efron and Hastie (2016) signify that computational statistics research can fall into two camps. It is generally possible to make a categorisation between algorithm development and algorithm evaluation, although this division is not fully impermeable. Within the statistical field, algorithmic development typically takes place with inferential operating characteristics and principles in mind. New algorithms for data analysis are also developed by a wide range of subject disciplines, often shaped using other criteria. Teasing out the underlying statistical context and significance of these innovative approaches can lead to theoretical insights and improved techniques. The statistical toolbox can be applied to algorithms in two different ways

- Development of new algorithms for statistical inference
- Evaluation of existing algorithms using statistical theory and methods.

Big Data hoists a big tent. Research contributions in the domain will come from many fields. Statistical contributions will inevitably take both of the aforementioned forms. We can take new algorithms to Big Data or we can seek to understand the statistical basis of promising methodology that has been proposed outside of the area. The work in this thesis falls into both of these categories, to be outlined in more detail shortly.

To locate our focus within the broader field of statistics it is effective to consider Box's loop, a conceptual process model of scientific data analysis (Box, 1976). Box's loop represents the cycle of model building, fitting, checking and refining that typically takes place when interacting with data, diagrammed in Figure 1.1. Following Blei (2014) it is possible to identify four key steps in Box's loop: Build, compute, critique and repeat. The initial build step is to postulate a plausible data generating process behind the observed data. Examples include but are certainly not limited to linear regression models, generalised

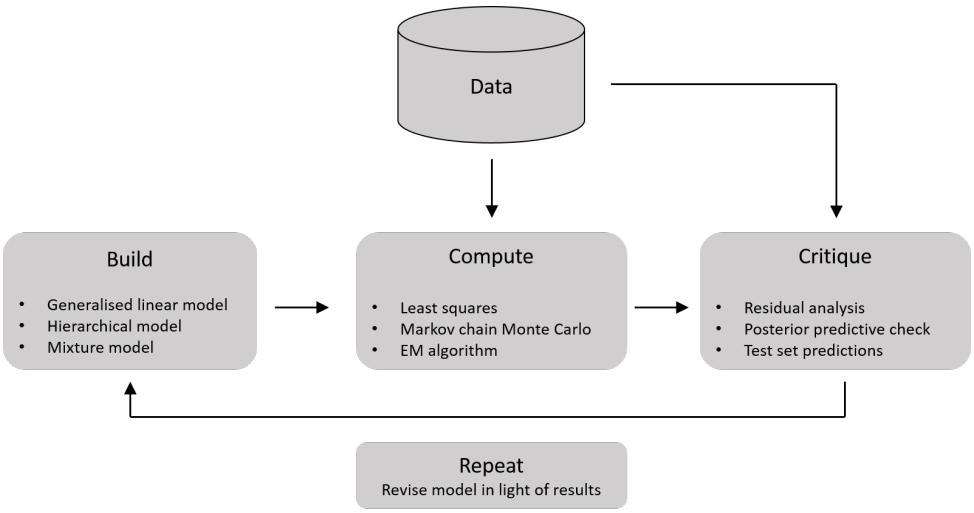


Figure 1.1: Box's loop is conceptual process model of scientific data analysis (Box, 1976). Box's loop defines a number of key phases (Build, compute, critique and repeat) when approaching a data modelling problem. The compartmentalisation of the overall task highlights important tactical decisions that are involved the statistical modelling lifecycle and aids the elicitation of broader strategic elements that influence the work. Adapted from Blei (2014).

linear models or mixture models. Substantial domain knowledge can be encoded at this stage. The second step is to compute the structural unknowns in the built model. This can involve the estimation of parameters through optimisation, or sampling from posterior distributions. The third step is to criticise the model through a formal procedure. Examples include residual diagnostics and posterior predictive checks. After the assessment, we then typically improve aspects of the model and repeat the process until we are satisfied. The final model is then deployed for prediction, visualisation, inference or any other suitable objective.

In an ideal world of infinite computational resources, the compute step would be instantaneous and painless. In reality we feel the pinch of scarcity, and with tall datasets the compute step can be a significant bottleneck. The work in this thesis is focused around the compute step of Box's loop. In particular, we are concerned with the development of scalable computational methods for Bayesian regression modelling on large datasets. Consider the general scenario where we have a dataset $\boldsymbol{y}$ of n observations with likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$ and prior $p(\boldsymbol{\theta})$, for $\boldsymbol{\theta} \in \Omega \subset \mathbb{R}^d$. Many of the key Monte Carlo methods for Bayesian computation require a full likelihood evaluation $p(\boldsymbol{y}|\boldsymbol{\theta})$ per loop (Robert, 2007; Robert and Casella, 2010). This a highly undesirable trait in the Big Data era when the $O(n)$ cost of likelihood evaluations can be substantial. We propose a number of computationally efficient algorithms for Bayesian regression modelling with tall (large n) datasets that address the likelihood burden. We focus on issues surrounding model selection, in particular computation of the integrated likelihood $p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$. The latter half of the thesis takes a particular focus on the Gaussian linear model, and how random projection can be used for computationally efficient statistical inference. In particular we illuminate some statistical properties of randomised data compression algorithms proposed in the computer science and machine learning literature.

The integrated likelihood $p(\boldsymbol{y})$ is also commonly known as the model evidence, and the quantity plays an important role in Bayesian model choice. We develop theory and methods for efficient computation of the integrated likelihood using parallel processing in Chapter 2. Chapter 3 explores how subsampling can be used for efficient estimation of the integrated likelihood. The parallel processing and subsampling algorithms are compatible with a variety of regression models. We consider their application to logistic regression in detail. Chapters 4, 5 investigate the use of random projection for approximate computation of the least squares estimates for Gaussian linear models. Chapter 6 extends the approach for carrying out approximate Bayesian model selection.

There have been many recent advances in developing scalable computational methods for Bayesian

inference in the fixed model setting. With the model given and fixed, the typical goal is to generate samples from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Uncertainty on $\boldsymbol{\theta}$ can shrink to the point of no practical significance in an analysis with a single model and huge n. This phenomenon is an instigative dynamic for our purposes. Model uncertainty as opposed to parameter uncertainty can often emerge as the more pressing inferential question given a massive number of observations (Varian, 2014). Bayesian model selection has received comparatively little attention in the Big Data literature, and much of the novelty in this thesis comes from the focus on this area.

We explore how distributed computing, subsampling and random projection can be used as methods for efficient Bayesian model selection. These techniques have been integrated into algorithms for sampling from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ when n is large. There are largely independent streams of literature surrounding each technique, Bardenet et al. (2017) provide an excellent survey of prior work. The problems of posterior sampling and integrated likelihood evaluation are connected, but not fully equivalent. Algorithms for the calculating the model evidence typically require the generation of posterior samples as an initial step. As algorithms for posterior simulation do not typically produce the model evidence as an ancillary benefit, the posterior samples must then be used in a secondary estimation procedure to obtain the model evidence (Friel and Wyse, 2012; Raftery, 1995; Skilling et al., 2006). The two-stage nature of the problem poses new research questions in the Big Data setting. Accelerated procedures for posterior sampling do not naturally lead to accelerated procedures for calculation of the integrated likelihood. Our use of parallel processing, subsampling and random projection for calculation of the model evidence has major differences compared to prior work that is focused on posterior simulation.

Distributed computing platforms lend themselves naturally to a divide and conquer approach for the analysis of tall datasets. The difficult Big Data job can be broken down into a series of smaller manageable tasks by splitting the full dataset into subsets, and analysing each subset on a separate machine. The subset results are then merged together to provide the end result. An important factor in a divide and conquer algorithm is the degree of communication required between subprocessors during the minibatch analyses. Embarrassingly parallel algorithms require no communication during this stage, and the only pooling of information occurs in a single round at the end. Embarrassingly parallel algorithms are attractive as not all distributed computing platforms allow for communication between individual workers. Secondly, the latency cost of regular communication between machines can be very high (Scott et al., 2013).

Subsampling also has a natural appeal for large n problems. It is sometimes possible to modify standard algorithms that involve full likelihood evaluations to instead use estimated likelihoods from a subsample of size m. Unsurprisingly, the efficiency of the modified algorithm depends on the quality of the likelihood estimator. A common finding in prior work is that Monte Carlo variance reduction techniques are needed to prevent algorithms from breaking down as the subsampling fraction m/n tends to zero (Maclaurin and Adams, 2014; Bierkens et al., 2016; Baker et al., 2017).

The computational demands of tall datasets can be addressed by sacrificing some accuracy in order to lower the computational expense of the analysis. Algorithms that play this trade-off to good effect have been identified as a promising future direction for Bayesian computation (Green et al., 2015). This principle leads us to consider 'sketching', a probabilistic data compression technique that has been developed largely within the computer science community. Sketching algorithms use random projection to generate a compressed dataset of k observations from the original source dataset of n observations. The compressed dataset is then used for approximate inference. Crucially, sketching algorithms offer stronger probabilistic guarantees on the stochastic approximation error that can be achieved through simple random sampling from the full dataset. Sketching algorithms have also shown excellent empirical performance compared to basic subsampling schemes in a number of simulation studies (Mahoney, 2011; Ma et al., 2015). Sketching has recently been explored as a method for approximate Bayesian inference in the fixed model setting (Geppert et al., 2017).

In Chapter 2 we develop an embarrassingly parallel algorithm for computation of integrated likelihood.

Data augmentation and Gibbs sampling are the key components of the method. In Chapter 3 we propose a computationally efficient method for estimating the integrated likelihood using subsampling. We find that it is difficult to modify existing importance sampling algorithms to use subsampling effectively. We propose an alternative interval estimator of the model evidence that has more favourable properties. As in related work, it is necessary to use variance reduction techniques to ensure the algorithm is stable even as the subsampling fraction m/n tends to zero.

Chapter 4 develops statistical properties of sketched regression algorithms in the fixed model setting. Much of the existing literature on sketching is from an algorithmic point of view, and we place sketching in a statistical context. This groundwork is needed before considering sketched model selection. Chapter 5 gives proofs for many of the results in Chapter 4. Proofs are presented separately in Chapter 5 so as to not hamper the exposition in Chapter 4. Of particular note is a proof of a central limit theorem where we show the sketched dataset has a matrix normal distribution under mild regularity conditions. The regularity conditions have an appealing interpretation in terms of the geometry of the source dataset. Chapter 6 explores the use of sketching for approximate Bayesian subset selection. Random projection can be used to approximate the integrated likelihood $p(\boldsymbol{y})$ in less time than is needed for exact computation. We find that the Tracy-Widom law (Tracy and Widom, 1994) is very useful for quantifying the error in the approximate integrated likelihood. The Tracy-Widom law describes the asymptotic distribution of the eigenvalues of large random matrices and has found many applications in high-dimensional statistics (Johnstone, 2006). The connection to sketching algorithms appears to be a new result.

Each chapter is largely self-contained and includes a review of relevant literature. Returning to the development/evaluation classification that was mentioned earlier, chapters 2 and 3 can be viewed as algorithm development. We propose new methodology for Bayesian computation with tall datasets. Chapters 4, 5 and 6 can be classified as algorithm evaluation. We investigate the statistical properties and principles that underpin existing sketching algorithms. As well of being of theoretical interest, this helps to define concrete procedures for assessing the quality of randomised algorithms.

Multiple real datasets are analysed throughout in order to demonstrate the theory and methods. Chapters 2 and Chapter 3 consider a number of benchmark datasets from the computational statistics literature. Chapters 4 and 6 analyse some large genetic datasets from the UK Biobank database. The dissertation concludes in Chapter 7 with a short discussion of the major lessons and common themes that have materialised over the enclosed work.

2

Split, apply, combine: computing the model evidence using embarrassingly parallel processing

Summary

We investigate how parallel processing can be used for computing the integrated likelihood on datasets with a large number of observations n. Tall datasets can be split into many subsets that are then allocated to different machines on a compute cluster. The subsets can then be analysed concurrently; embarrassingly parallel algorithms run each minibatch analysis with no cross communication between subprocessors. We find that a combination of data augmentation and Gibbs sampling facilitates a simulation consistent and scalable embarrassingly parallel algorithm for a wide class of statistical models. We show that conditionally conjugate exponential family models exhibit structure that is amenable to embarrassingly parallel inference. A second theoretical finding involves the final step of the divide and conquer algorithm where the subset results are pooled to give the final estimate of the model evidence. We show this step has an interpretation as a Bayesian hypothesis testing problem. The connection furthers the theory of distributed Bayesian inference and leads to a second simulation consistent algorithm for computing the model evidence in parallel.

2.1 Introduction

Distributed computing platforms offer a huge amount of computing power that can be harnessed to complete many tasks simultaneously. Using parallel processing for the statistical analysis of large datasets is thus an appealing idea. The broad paradigm is to first *split* the full dataset into a collection of non-overlapping subsets. Subsets are then allocated to different machines on a network, we then *apply* conventional statistical methods to each minibatch of data. The final step is to *combine* the subset output into a single estimate. Separating the algorithm into distinct *split*, *apply* and *combine* phases is useful for algorithm design and evaluation. Figure 2.1 displays a broach schematic of the desired approach. A stringent desiderata of an embarrassingly parallel scheme is that the *combine* phase does not allow interaction with the original dataset. The entire dataset is processed and distilled in the *apply* phase. Embarrassingly parallel algorithms follow a divide and conquer principle, where the original difficult problem is broken down into a series of manageable tasks.

This computational blueprint presents novel challenges for statistical inference. It can be difficult to identify appropriate summary measures for the subset analyses conducted in the *apply* stage, and the final aggregation in the *combine* phase is a challenging evidence synthesis task (Jordan, 2013). These issues are present when implementing a divide and conquer approach for Bayesian model selection. Typically, each subset will be analysed using Markov chain Monte Carlo methods and will return a rich set of output. An additional consideration with model selection is that individual subset analyses will be underpowered relative to a full dataset analysis. We are unlikely to see support for complex models in the subset analyses, despite the fact that we may have a large enough dataset to detect complex features. We

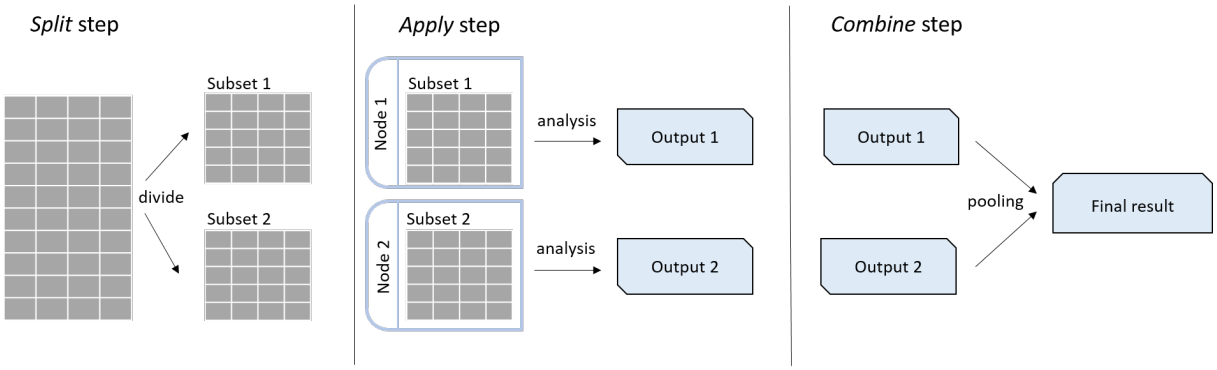


Figure 2.1: Template for embarrassingly parallel algorithms. The split step breaks the full dataset in to s non overlapping subsets. The illustration is for $s = 2$. Each subset is then allocated to a different machine on a cluster. During the *apply* step we apply conventional methodology to each data subset with no cross communication between workers. Each analysis is summarised by a consistent format of output. The s sets of output from the apply stage are then synthesised in the combine step to give the final result. In this design brief, the combine stage only involves the output from the apply step, and not the original dataset.

explore theoretical and practical issues for distributed Bayesian model selection. Of particular concern is how to summarise the subset analyses effectively in the apply step and how to synthesise the minibatch results appropriately into an overall consensus in the combine step.

Suppose interest lies in a collection of M parametric models $\mathcal{M}_1, \ldots, \mathcal{M}_M$. The prior probability of model j is given by $p(\mathcal{M}_j)$ for $j = 1, \ldots, M$. In a slight abuse of notation we do not write $\boldsymbol{\theta}_j$ to specify the parameter associated with model $\mathcal{M}_j$, this is to stop notation from being cluttered. The notation $p(\boldsymbol{\theta}|\mathcal{M}_j)$ is implicitly understood to refer to the distinct parameter associated with model $\mathcal{M}_j$. In this work we assume that model selection is to be carried out by calculating the integrated likelihood $p(\boldsymbol{y}|\mathcal{M}_j)$ for each model $j = 1, \ldots, M$. The posterior distribution over models can be calculated easily by enumerating over the full set of M integrated likelihoods and prior probabilities:

$$p(\mathcal{M}_j|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\mathcal{M}_j)p(\mathcal{M}_j)}{\sum_{k=1}^{M} p(\boldsymbol{y}|\mathcal{M}_k)p(\mathcal{M}_k)}. \tag{2.1}$$

It is therefore sufficient to consider the distributed computation of the integrated likelihood for some arbitrary model $\mathcal{M}$. Given an algorithm for an arbitrary model, we simply apply it to each model in the set of interest and then enumerate to obtain the posterior probabilities using (2.1). Assume the parameter space of model $\mathcal{M}$ is indexed by some continuous $\boldsymbol{\theta} \in \Omega \subset \mathbb{R}^d$. Let the data $\boldsymbol{y}$ consist of n independent observations. Given a prior distribution $p(\boldsymbol{\theta})$ and a likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$, the integrated likelihood is defined as

$$p(\boldsymbol{y}|\mathcal{M}) = \int p(\boldsymbol{y}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})\ d\boldsymbol{\theta}. \tag{2.2}$$

Previous work on Bayesian divide and conquer algorithms is largely focused on how to sample from or approximate the target posterior $p(\boldsymbol{\theta}|\boldsymbol{y}, \mathcal{M})$ (Huang and Gelman, 2005; Scott et al., 2013; Neiswanger et al., 2013; Srivastava et al., 2015). Calculation of the integrated likelihood (2.2) presents different challenges, and we largely build on the existing literature for computing integrated likelihoods and Bayes factors in the single machine setting. The first main theoretical finding is that the combine step in the divide and conquer algorithm has an interpretation as a Bayesian hypothesis testing problem. The gold standard Bayesian evidence synthesis rule can be expressed as a Savage-Dickey density ratio (Dickey, 1971). We also find that data augmentation (Tanner and Wong, 1987) and the theory of conditional conjugacy for exponential family is of use to define appropriate action in the apply and combine stages.

From a computational point of view, we propose two simulation consistent algorithms for calculating the model evidence in parallel. The first algorithm is developed in section 2.4 and is designed for general parametric models. The algorithm presumes that the apply stage involves a generic Metropolis-Hastings

sampler run on each subset. The subset output consists of the posterior draws. The combine phase then consists of a density estimation task given the subset posterior samples. The dimension of the density estimation task increases linearly with the number of sub-processors, thus limiting its scalability. The second algorithm is developed in section 2.5, and applies to conditionally conjugate exponential family models. We assume that conjugacy can be achieved through a suitable data augmentation scheme. The algorithm prescribes Gibbs sampling in the apply step. The output is no longer the raw sampled values, but rather the parameters of the conditional posterior at each stage of the Gibbs run. The combine stage then involves aggregation of the Gibbs sampler histories from each subset. Data augmentation proves useful as we can then obtain closed form rules for combining the output from the apply stage. The approach is related to the method of Chib (1995) for computing the integrated likelihood from the Gibbs output. Although limited in scope compared to the first algorithm, the approach involving data augmentation has significantly more favourable computational properties. By combining data augmentation and Gibbs sampling with parallel processing, it is possible to conduct Bayesian model selection using a divide and conquer strategy. We illustrate the theory and methods on a number of real datasets, with a particular focus on logistic regression.

2.2 Divide and conquer Bayesian inference

2.2.1 Posterior sampling

With a specified model, the divide and conquer approach is largely motivated by a factorisation of the full dataset posterior distribution into a combination of subset posterior distributions (Bardenet et al., 2017). We will assume it is possible to partition the data into s subsets $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)$ such that the subsets are independent given $\boldsymbol{\theta}$. To describe the strategy we introduce the idea of a subprior distribution and a subposterior distribution. The subprior distribution $\widetilde{p}(\boldsymbol{\theta})$ is defined as

$$\widetilde{p}(\boldsymbol{\theta}) = \frac{p(\boldsymbol{\theta})^{1/s}}{\int p(\boldsymbol{\theta})^{1/s}\, d\boldsymbol{\theta}}.$$

We assume that the fractionated prior $p(\boldsymbol{\theta})^{1/s}$ is integrable so that the subprior distribution is well defined. The assumption is always satisfied for Gaussin priors, however in general this condition needs to be checked on case by case basis. The subprior distribution contains a fraction of the prior information encoded in the original prior distribution $\widetilde{p}(\boldsymbol{\theta})$. The subposterior distributions are defined in terms of the subset likelihoods and the subprior distributions. For $i = 1, \ldots, s$ the i-th subposterior distribution $\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i)$ is defined as

$$\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i) = \frac{p(\boldsymbol{y}_i|\boldsymbol{\theta})\widetilde{p}(\boldsymbol{\theta})}{\int p(\boldsymbol{y}_i|\boldsymbol{\theta})\widetilde{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta}}.$$

The tilde is used to acknowledge the use of the subprior distribution $\widetilde{p}(\boldsymbol{\theta})$ in place of the original prior distribution $p(\boldsymbol{\theta})$. All probability formulae conditional on $\boldsymbol{\theta}$ do not require a tilde, so we use the regular notation for the subset likelihood $p(\boldsymbol{y}_i|\boldsymbol{\theta})$. We define the subprior normalising constant α as $\alpha = \int p(\boldsymbol{\theta})^{1/s}\, d\boldsymbol{\theta}$ and the subposterior evidence as $\widetilde{p}(\boldsymbol{y}_i) = \int p(\boldsymbol{y}_i|\boldsymbol{\theta})\widetilde{p}(\boldsymbol{\theta})\, d\boldsymbol{\theta}$. Table 2.1 lists some key terms that we will make use of when discussing divide and conquer Bayesian inference.

A divide and conquer strategy can be motivated by noting that full dataset posterior is proportional

Term	Symbol	Definition
Subprior	$\widetilde{p}(\boldsymbol{\theta})$	$\widetilde{p}(\boldsymbol{\theta}) = p(\boldsymbol{\theta})^{1/s}/\alpha$.
Subposterior	$\widetilde{p}(\boldsymbol{\theta}\|\boldsymbol{y_i})$	$\widetilde{p}(\boldsymbol{\theta}\|\boldsymbol{y_i}) = p(\boldsymbol{y}_i\|\boldsymbol{\theta})\widetilde{p}(\boldsymbol{\theta})/\widetilde{p}(\boldsymbol{y}_i)$.
Subprior normalising constant	α	$\alpha = \int p(\boldsymbol{\theta})^{1/s}\ d\boldsymbol{\theta}$.
Suposterior evidence	$\widetilde{p}(\boldsymbol{y}_i)$	$\widetilde{p}(\boldsymbol{y}_i) = \int p(\boldsymbol{y}_i\|\boldsymbol{\theta})\widetilde{p}(\boldsymbol{\theta})\ d\boldsymbol{\theta}$.

Table 2.1: Terminology for divide and conquer Bayesian inference. The notation and terminology in this table is implicitly conditonal on some model $\mathcal{M}$ with parameter $\boldsymbol{\theta}$.

to the product of s subposterior distributions.

$$
\begin{aligned}
p(\boldsymbol{\theta}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s) &= \frac{p(\boldsymbol{\theta})p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\boldsymbol{\theta})}{p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)} \\
&= \frac{1}{p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)} p(\boldsymbol{\theta}) \prod_{i=1}^{s} p(\boldsymbol{y}_i|\boldsymbol{\theta}) \\
&= \frac{1}{p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)} \prod_{i=1}^{s} p(\boldsymbol{y}_i|\boldsymbol{\theta}) p(\boldsymbol{\theta})^{1/s} \\
&= \frac{\alpha^s}{p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)} \prod_{i=1}^{s} p(\boldsymbol{y}_i|\boldsymbol{\theta}) \widetilde{p}(\boldsymbol{\theta}) \\
&= \frac{\alpha^s \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{y}_i)}{p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)} \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i) \qquad (2.3) \\
&\propto \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i). \qquad (2.4)
\end{aligned}
$$

Dropping the constants in (2.3) leads to the relationship in (2.4). Each data subset $\boldsymbol{y}_i$ can be distributed to a subprocessor. It is then possible to generate samples from each subposterior $\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i)$ in parallel, typically by running MCMC on each worker. When posterior simulation is the ultimate goal, the combine step must pool the subposterior samples in a manner such that the full dataset posterior $p(\boldsymbol{\theta}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)$ is targeted.

A variety of combination rules have been suggested for synthesising the subposterior output in order to target $p(\boldsymbol{\theta}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)$. Huang and Gelman (2005) and Scott et al. (2013) develop rules based on making a normal approximation to each subposterior. Neiswanger et al. (2013) propose an approach using kernel density estimation. Wang and Dunson (2013) consider the use of the Weierstrass transform. There are also aggregation rules based on geometric or robustness arguments (Srivastava et al., 2015; Minsker et al., 2017). As our goal is calculation of the integrated likelihood as opposed to posterior sampling, we do not review any of these methods in detail. Divide and conquer Bayesian model selection presents different challenges as the decomposition of the full data model posterior into batch posteriors is more complicated than in (2.4).

2.2.2 Model uncertainty

Motivating a divide and conquer approach for model selection requires a different line of reasoning compared to the fixed model case, as the full dataset integrated likelihood is not equal to the product of the subset integrated likelihoods,

$$
\begin{aligned}
p(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_k|\mathcal{M}_j) &= p(\boldsymbol{y}_1|\mathcal{M}_j) \prod_{i=2}^{s} p(\boldsymbol{y}_i|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{i-1}, \mathcal{M}_j) \\
&\neq \prod_{i=1}^{s} p(\boldsymbol{y}_i|\mathcal{M}_j).
\end{aligned}
$$

The first line is simply the chain rule of probability. The inequality is present as the subsets are only conditionally independent when given both the model and the parameters. For example, if the underlying model is a normal distribution, the observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s$ are still dependent until the mean and variance parameters are completely specified. Although we have assumed $p(\boldsymbol{y}_i|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{i-1}, \boldsymbol{\theta}, \mathcal{M}_j) = p(\boldsymbol{y}_i|\boldsymbol{\theta}, \mathcal{M}_j)$, the integrated likelihood does not have the same property $p(\boldsymbol{y}_i|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{i-1}, \mathcal{M}_j) \neq p(\boldsymbol{y}_i|\mathcal{M}_j)$. As such, we cannot immediately obtain a subposterior factorisation over data batches using the same argument in (2.4).

The target integrated likelihood is related to the subposterior evidence values, the subprior normalising constant and the subposterior distributions. Rearranging (2.3) gives

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mathcal{M}_j)p(\boldsymbol{\theta}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mathcal{M}_j) = \alpha^s \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{y}_i|\mathcal{M}_j) \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \mathcal{M}_j).$$

Integrating both sides over $\boldsymbol{\theta}$ gives an important expression for the full dataset model evidence.

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mathcal{M}_j) = \alpha^s \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{y}_i|\mathcal{M}_j) \int \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \mathcal{M}_j)\ d\boldsymbol{\theta}. \tag{2.5}$$

The full dataset model evidence is related to the subposterior evidence values, the subprior normalising constant α and an integral over the subposterior distributions. The complex relationship between the full dataset evidence and the subset output means that is is difficult to obtain a simple rule for embarrassingly parallel model selection. Define the subprior model probabilities as

$$\widetilde{p}(\mathcal{M}_j) = \frac{p(\mathcal{M}_j)^{1/s}}{\sum_{k=1}^{M} p(\mathcal{M}_k)^{1/s}}.$$

We define the subposterior model probabilities $\widetilde{p}(\mathcal{M}_j)$ in terms of the subposterior evidence $\widetilde{p}(\boldsymbol{y}_i|\mathcal{M}_j)$ and the subprior model probability $\widetilde{p}(\mathcal{M}_j)$:

$$\widetilde{p}(\mathcal{M}_j|\boldsymbol{y}_i) = \frac{\widetilde{p}(\boldsymbol{y}_i|\mathcal{M}_j)\widetilde{p}(\mathcal{M}_j)}{\sum_{k=1}^{M} \widetilde{p}(\boldsymbol{y}_i|\mathcal{M}_k)\widetilde{p}(\mathcal{M}_k)}. \tag{2.6}$$

It is more difficult represent the full dataset posterior over models as a combination of subposterior distributions. We use α_j to denote the subprior normalising constant for model $\mathcal{M}_j$, so $\alpha_j = \int p(\boldsymbol{\theta}|\mathcal{M}_j)^{1/s}\ d\boldsymbol{\theta}$.

The full dataset posterior is related to the subposterior model probabilities, the subprior normalising constants and the subposterior distributions on the model parameters. We start by writing out the decomposition in full.

$$\begin{aligned} p(\mathcal{M}_j|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s) &= \frac{p(\mathcal{M}_j)p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mathcal{M}_j)}{\sum_{k=1}^{M} p(\mathcal{M}_k)p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mathcal{M}_k)} \\ &\propto p(\mathcal{M}_j)p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mathcal{M}_j). \end{aligned} \tag{2.7}$$

Now substituting in (2.5) and using the fact that $\widetilde{p}(\mathcal{M}_j) \propto p(\mathcal{M}_j)^{1/s}$,

$$\begin{aligned} p(\mathcal{M}_j|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s) &\propto p(\mathcal{M}_j) \left(\prod_{i=1}^{s} \widetilde{p}(\boldsymbol{y}_i|\mathcal{M}_j) \right) \alpha_j^s \int \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \mathcal{M}_j)\ d\boldsymbol{\theta} \\ &= \left(\prod_{i=1}^{s} \widetilde{p}(\boldsymbol{y}_i|\mathcal{M}_j)p(\mathcal{M}_j)^{1/s} \right) \alpha_j^s \int \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \mathcal{M}_j)\ d\boldsymbol{\theta} \\ &= \left(\prod_{i=1}^{s} \widetilde{p}(\boldsymbol{y}_i|\mathcal{M}_j)\widetilde{p}(\mathcal{M}_j) \right) \alpha_j^s \int \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \mathcal{M}_j)\ d\boldsymbol{\theta}. \end{aligned} \tag{2.8}$$

The following quantity is a constant as it involves sums over the collection of models

$$\prod_{i=1}^{s} \frac{1}{\sum_{k=1}^{M} \widetilde{p}(\boldsymbol{y}_i|\mathcal{M}_k)\widetilde{p}(\mathcal{M}_k)} \tag{2.9}$$

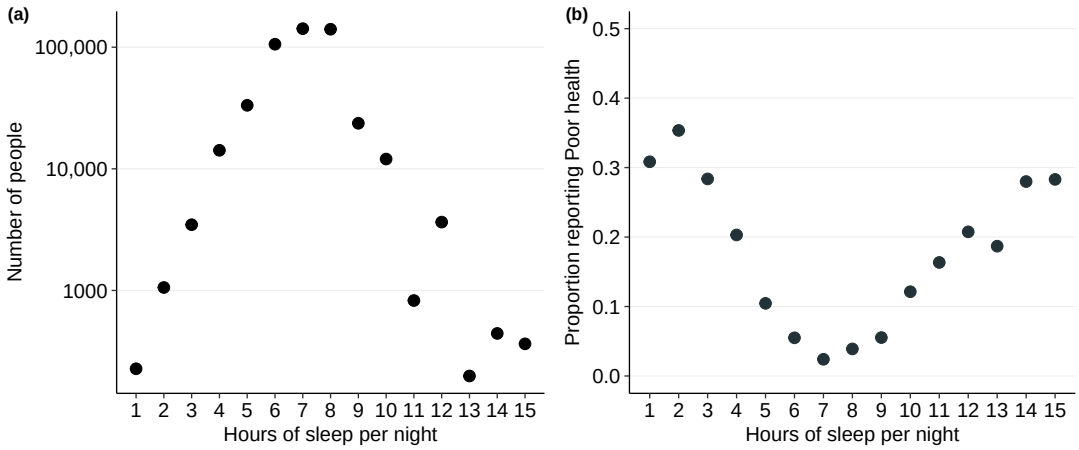


Figure 2.2: Sleep dataset from the Behavioural Risk Factor Surveillance survey ($n = 481,939$). Panel (a) shows the number of individuals reporting a particular number of hours sleep. Panel (b) shows the proportion of individuals reporting 'Poor' health against hours sleep. Self-reported 'Poor' health status appears to be statistically associated with sleep habits.

Multiplying (2.8) by the constant in (2.9) yields

$$p(\mathcal{M}_j|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s) \propto \left(\prod_{i=1}^{s} \frac{\widetilde{p}(\boldsymbol{y}_i|\mathcal{M}_j)\widetilde{p}(\mathcal{M}_j)}{\sum_{k=1}^{M}\widetilde{p}(\boldsymbol{y}_i|\mathcal{M}_k)\widetilde{p}(\mathcal{M}_k)}\right)\alpha_j^s \int \prod_{i=1}^{s}\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i,\mathcal{M}_j)\; d\boldsymbol{\theta}$$
$$= \left(\prod_{i=1}^{s}\widetilde{p}(\mathcal{M}_j|\boldsymbol{y}_i)\right)\alpha_j^s \int \prod_{i=1}^{s}\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i,\mathcal{M}_j)\; d\boldsymbol{\theta}. \tag{2.10}$$

The full dataset posterior can be expressed as

$$p(\mathcal{M}_j|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s) \propto \left(\prod_{i=1}^{s}\widetilde{p}(\mathcal{M}_j|\boldsymbol{y}_i)\right)\alpha_j^s \int \prod_{i=1}^{s}\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i,\mathcal{M}_j)\; d\boldsymbol{\theta}. \tag{2.11}$$

The immediate message from equation (2.11) is that the subposterior model probabilities $\widetilde{p}(\mathcal{M}_j|\boldsymbol{y}_1),\ldots,\widetilde{p}(\mathcal{M}_j|\boldsymbol{y}_s)$ are not sufficient to evaluate the global suitability of a model. The subposterior distributions of the parameters $\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_1,\mathcal{M}_j),\ldots,\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_s,\mathcal{M}_j)$ contain important additional information for discriminating between models.

To illustrate, we first consider a simple model choice problem. The dataset is from the 2013 Behavioural Risk Factor Surveillance System (BRFSS) survey run by the Centers for Disease Control and Prevention (Centers for Disease Control and Prevention, 2013). We examine the $n = 481,939$ responses for two general health questions. Respondents were asked for how many hours per night do they typically sleep. Respondents were also asked to rate their general health on a five point scale: Poor, Fair, Good, Very Good, Excellent. Figure 2.2 (a) plots the number of responses against hours sleep. The majority of those surveyed report between 6 and 8 hours sleep, with the overall counts decreasing as the hours of sleep moves away from this range. The y-axis is on a log scale. Figure 2.2 (b) plots the proportion of respondents reporting 'Poor' health against hours sleep. There appears to be a statistical association between the two variables. Atypical sleep habits, in terms of very low or very high hours of sleep appears to be an indicator for self-reported 'Poor' health.

As a simple model choice exercise we fit two different logistic regression models to the data in (b). Let x_i denote the hours of sleep for individual i, and y_i be a binary indicator for health status, where $y_i = 1$ if the reported health is 'Poor' and $y_i = 0$ otherwise. The response is modelled as $y_i \sim \text{Bernoulli}(\sigma(\eta_i))$ for a linear predictor η_i, where $\sigma(\eta)$ gives the inverse logistic function $\sigma(\eta) = 1/(1+\exp(-\eta))$. We compare a cubic model ($\mathcal{M}_1$) with four parameters to a more complex cubic spline model ($\mathcal{M}_2$) with 10 parameters. Under the simpler model $\mathcal{M}_1$, the linear predictor is given by

$$\mathcal{M}_1 : \eta_i = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + x_i^3\beta_3.$$

For the cubic spline model $\mathcal{M}_2$ with K knots we set

$$\mathcal{M}_2 : \eta_i = \beta_0 + x_1\beta_1 + x_i^2\beta_2 + x_i^3\beta_3 + \sum_{k=1}^{K}(x_i - \xi_k)_+^3$$

We chose $K = 6$ knots manually. The knots $\xi_1, \ldots, \xi_6$ were set at 3,5,7,9,11 and 13. As a brief aside, it worth acknowledging that the dataset is from a complex health survey and that the Bayesian models were fit under the assumption that the data collection mechanism is ignorable (Gelman et al., 2014).

We conducted a divide and conquer analysis by splitting the data into $s = 2$ subsets, based on the x_i values. Observations with $x_i \leq 7$ were placed in subset 1, and observations with $x_i > 7$ were allocated to subset 2. This partition was a deliberate choice to highlight the influence of the split step on the combine step. Figure 2.3 shows the fitted models on the full dataset and the subsets. In each panel we plot the posterior predictive mean function from the respective analysis as a red line. Using the full dataset, the posterior predictive mean for a new response given covariates $\boldsymbol{x}_{\text{new}}$ is obtained by integrating over the posterior distribution of the coefficients

$$\mathbb{E}[y_{\text{new}}|\mathcal{M}] = \int p(y_{\text{new}} = 1|x_{\text{new}}, \boldsymbol{\beta}, \mathcal{M})p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y}, \mathcal{M})\ d\boldsymbol{\beta}. \tag{2.12}$$

Let $\boldsymbol{X}_{(i)}$ denote the design matrix for data subset i for $i = 1, 2$. For the subset results, the posterior predictive mean for a new response given covariates $\boldsymbol{x}_{\text{new}}$ is is obtained by integrating over the subposterior distribution of the coefficients

$$\mathbb{E}[y_{\text{new}}|\mathcal{M}] = \int p(y_{\text{new}} = 1|x_{\text{new}}, \boldsymbol{\beta}, \mathcal{M})\widetilde{p}(\boldsymbol{\beta}|\boldsymbol{X}_{(i)}, \boldsymbol{y}_i, \mathcal{M})\ d\boldsymbol{\beta}. \tag{2.13}$$

The shaded ribbons in each plot give 90 percent credible intervals for the posterior predictive mean at a given point. As expected, there is more uncertainty in the regions where we have fewer data points. From panel (a) in Figure 2.2 the overwhelming majority of respondents report between 4 and 12 hours of sleep, and the shaded ribbon is tightly concentrated around the posterior predictive mean in this range.

Looking at the full dataset results in Figure 2.3 it is apparent that the spline model is the more appropriate model. The subset results are less definitive. Visually, the fits of the spline and cubic models look comparable in each subset. In particular, the posterior mean functions and credible intervals look nearly identical in subset 1. Table 2.2 reports the full dataset posterior model probabilities and the subposterior model probabilities. The full dataset posterior puts mass one on the cubic spline model, this is not surprising given that the cubic model fits the observed data very poorly in the tails of the design space (hours of sleep). The subset results do not reflect this. In subset 1, the cubic model provides a seemingly identical fit to the spline model using the available data. As the cubic model is more parsimonious, it is favoured in the subposterior distribution on models with $\widetilde{p}(\mathcal{M}_1|\boldsymbol{y}_1, \boldsymbol{X}_{(1)}) = 0.84$. In subset 2 the extra parameters of the cubic spline model give a better fit to the available data. As such, the spline model is strongly favoured in the second minibatch analysis. The subposterior probability for the cubic model is $\widetilde{p}(\mathcal{M}_1|\boldsymbol{y}_2, \boldsymbol{X}_{(2)})$ is zero in the second subset. It is difficult to intuitively reconcile the subset results with the full dataset results when only given subposterior and posterior model probabilities. We freely admit that this example has been artificially engineered to be melodramatic. We wished to demonstrate that the choice of partition in the split step has a large downstream effect on the difficulty in the combine step. Additionally, we would like to highlight the role of subposterior overlap as a goodness of fit diagnostic in the combine step.

As dictated by (2.11), to properly synthesise the subset results it is necessary to look at the subposterior distributions of the parameters. The most direct interpretation of the integral term is an expectation over an arbitrary subposterior:

$$\int \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \mathcal{M}_j)\ d\boldsymbol{\theta} = \mathbb{E}_{\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \mathcal{M}_j)}\left(\prod_{k\neq i}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_k, \mathcal{M}_j)\right). \tag{2.14}$$

	Posterior distribution	Subposterior 1	Subposterior 2
Cubic model $\mathcal{M}_1$	0.00	0.84	0.00
Spline model $\mathcal{M}_2$	1.00	0.16	1.00

Table 2.2: Posterior and subposterior model probabilities for the sleep dataset. The subposterior results are in conflict despite very clear results in a full dataset analysis.

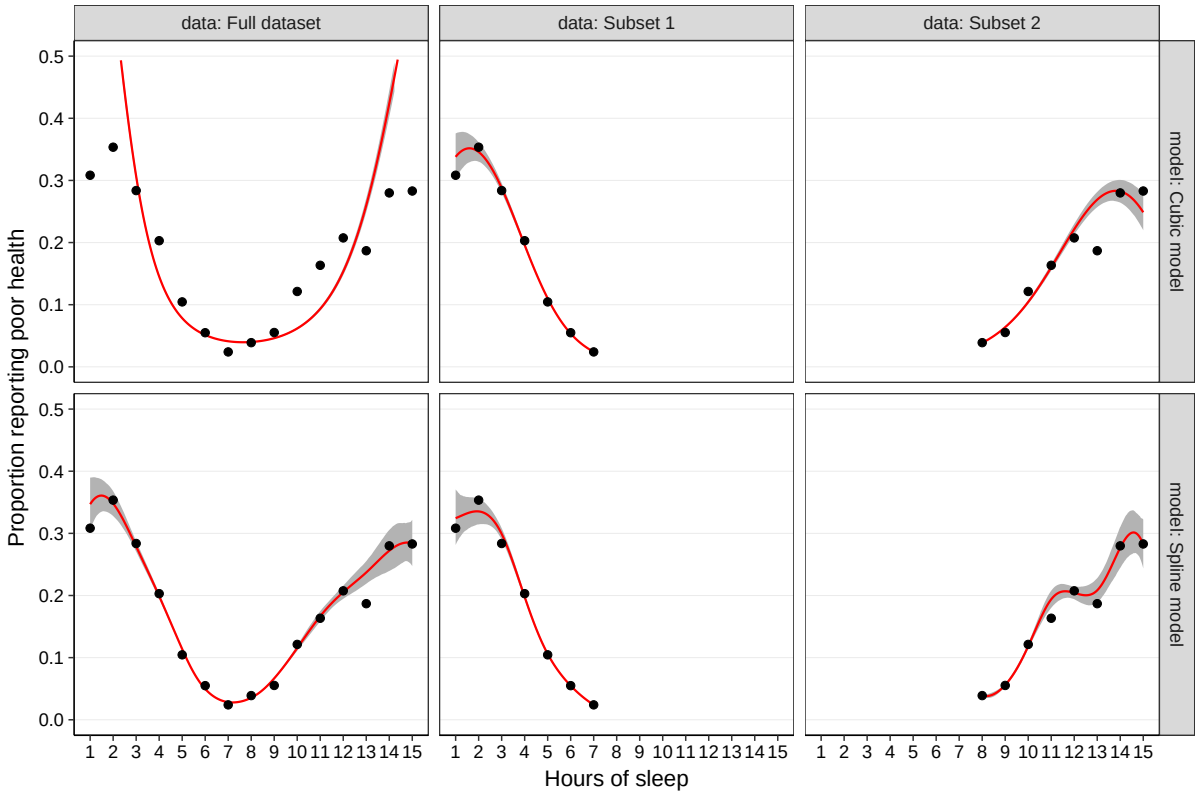


Figure 2.3: Comparison of cubic model and spline model on the full sleep dataset and on the $s = 2$ subsets. The red line gives the posterior preditive mean function. The shaded ribbons give a 90 percent credible interval for the posterior predictive mean. The full dataset contains $n = 481,939$ observations. Subset 1 contains $n_1 = 300,177$ observations, and subset 2 contains $n_2 = 181,762$ observations. The full dataset results show that the spline model provides a much better fit to the observed data. The cubic model does not fit well in the extreme regions of feature space (hours of sleep). The subset results in isolation do not make this as clear. Model 1 and Model 2 given nearly identical fits in subset 1. The results seem comparable in subset 2. The results in the apply stage may not be representative of a single full dataset analysis. The initial partition of the dataset in the split stage appears to influence the difficulty of the evidence synthesis task in the combine stage.

Qualitatively speaking, the expectation will be large if the subposterior distributions are similar across subsets, and the expectation will be small if the subposterior distributions are dissimilar. A high-level interpretation is that the integral checks for consistency of parameter estimates across different subsets. Figure 2.4 compares the marginal subposterior distributions on each parameter in the cubic model. The non-overlapping subposteriors are indicative of poor goodness of fit, which is clear in the full dataset results in Figure 2.3.

The example shows that the subposterior integral (2.14) has an important role in divide and conquer model selection. The subset model probabilities in isolation do not give enough information to reconstruct the target posterior distribution over models. Distributed model selection requires a pooling rule in the combine stage that takes this into account. It is quite straightforward to make an approximation to the subposterior integral, under the assumption of Gaussian subposteriors. Normal approximations to the subposterior distributions are used in the fixed model divide and conquer work by Huang and Gelman

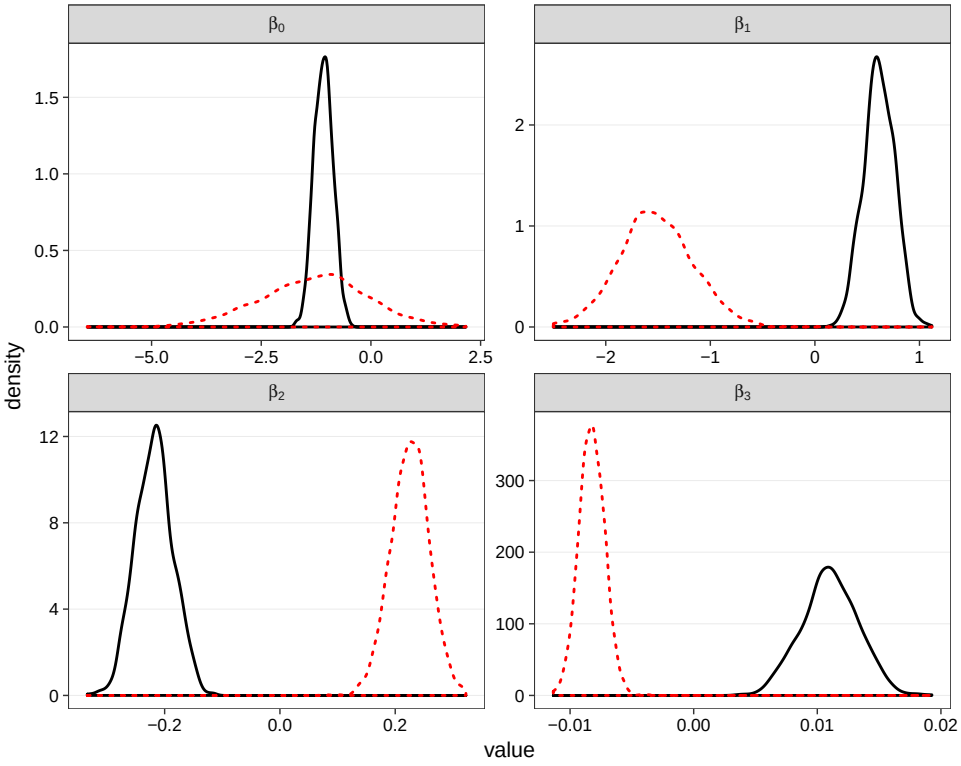


Figure 2.4: Suposterior distributions of parameters in the cubic model. The solid line denotes the first subposterior $\widetilde{p}(\boldsymbol{\beta}|\boldsymbol{y}_1)$. The red dashed line denotes the second subposterior $\widetilde{p}(\boldsymbol{\beta}|\boldsymbol{y}_2)$. The disparate subposterior distributions is a reflection of the poor overall fit of the model. Subposterior overlap is an important diagnostic in the combine stage for model selection.

(2005) and Scott et al. (2013). However, it is of interest to identify general strategies that do not rely on subposterior normality, as quantifying the error from making normal approximations can be difficult (Bierkens et al., 2016). Furthermore, a general solution can lead to a deeper understanding of the problem.

The combine step in divide and conquer model selection involves estimation of subposterior integral given the output from the apply stage. For later reference we denote the subposterior integral as I_{sub} where

$$I_{\text{sub}} = \int \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \mathcal{M}) \; d\boldsymbol{\theta}. \tag{2.15}$$

In section 2.4 we establish an important link between the subposterior integral and the Savage-Dickey density ratio. The connection explains how the subposterior integral measures subposterior overlap from a Bayesian perspective, and lays out how the combine step implicitly synthesises the support for a model given the output from the apply stage. The theoretical connection also leads to simple Monte Carlo procedure for estimating the integral involving kernel density estimation. Although simulation consistent, the estimator scales poorly with the number of subsets s as it involves kernel density estimation in high-dimensional spaces. To remedy this problem we propose an alternative strategy using data augmentation and Gibbs sampling in section 2.5. Although limited to certain models, the second algorithm gives a significantly better estimator of the subposterior integral (2.15) in the combine step. Section 2.3 reviews some important background theory.

Dimension	Sample size
1	4
2	19
3	67
4	223
5	678
6	2,790
7	10,700
8	43,700
9	187,000
10	842,000

Table 2.3: Required sample size to attain a relative mean square error of less than 0.1 using nonparametric kernel density estimation against dimension. The generative model is a d-dimensional multivariate normal distribution with an identity covariance matrix. The density is estimated at zero using a Gaussian kernel. The bandwidth is chosen to minimise the asymptotic mean square error at zero. The required sample size increases rapidly with the dimension d. Reproduced from Silverman (1986, Table 4.2)

2.3 Background

2.3.1 Integrated likelihood calculation

Estimation of the integrated likelihood has been a long standing problem in Bayesian computation. Rearranging Bayes' theorem, an important general relationship is that for any ordinate $\boldsymbol{\theta} \in \Omega$,

$$\log p(\boldsymbol{y}|\mathcal{M}) = \log p(\boldsymbol{y}|\boldsymbol{\theta}, \mathcal{M}) + \log p(\boldsymbol{\theta}|\mathcal{M}) - \log p(\boldsymbol{\theta}|\boldsymbol{y}, \mathcal{M}). \tag{2.16}$$

The identity (2.16) is often referred to as 'Candidate's formula' (Besag, 1989). If the posterior density can be obtained at the point $\boldsymbol{\theta}$ then we have a simple way to obtain the integrated likelihood. Given samples from the posterior distribution, it is possible to use a non-parametric kernel density estimator to estimate the density at $\boldsymbol{\theta}$ (Raftery, 1996; Lewis and Raftery, 1997). A problem with this approach is that kernel density estimation is subject to the curse of dimensionality, with an asymptotically optimal pointwise mean squared error rate of $O(B^{-4/(d+4)})$, where B is the number of independent posterior samples (Terrell and Scott, 1992). The bound is under the assumption that we use a second order kernel. Most common kernels (eg. Gaussian, Epanechnikov, Uniform) are second order kernels (Terrell and Scott, 1992). Use of kernel density estimation is therefore limited to models with a very small number of parameters. The impact of the curse of dimensionality is vividly presented by Silverman (1986, section 4.5.2). Silverman considers nonparametric kernel density estimation where the true density f is a d-dimensional multivariate Gaussian with identity covariance matrix. The task is to estimate the true density at zero $f(\mathbf{0})$ using a Gaussian kernel. Silverman sets the single bandwidth parameter h to minimise the asymptotic pointwise mean squared error. Suppose the goal is to ensure the relative mean square error $\mathbb{E}[(\widehat{f}(\mathbf{0}) - f(\mathbf{0}))^2]/[f(\mathbf{0})^2]$ is less than 0.1. Silverman calculates the required sample size to attain this error tolerance at various dimensions d. Table 2.3 reports the required sample size against dimensionality. The required sample size increases rapidly with d. Nonparametric kernel density estimation is very hard in high-dimensional spaces.

2.3.2 Savage-Dickey density ratio

The Savage-Dickey density ratio is a useful identity for computing Bayes factors that avoids the need to compute integrated likelihoods (Dickey, 1971) directly. For our purposes, we are concerned with the use of the Savage-Dickey density ratio to linear restrictions on parameters. The Savage-Dickey density

ratio for testing linear restrictions is discussed Wetzels et al. (2010) and McCulloch and Rossi (1992). We present the approach as in McCulloch and Rossi (1992). The Savage-Dickey identity expresses Bayes factors as a density ratio involving posterior and prior distributions. Suppose we have a model $\mathcal{M}_S$ with parameter $\boldsymbol{\theta}_S \in \Omega_S$. Let $L(\cdot) = f(\boldsymbol{y}|\cdot)$ denote the likelihood function, where the data $\boldsymbol{y}$ is treated as fixed. Let $\boldsymbol{\psi}$ represent a linear function of $\boldsymbol{\theta}_S$ so $\boldsymbol{\psi} = \boldsymbol{A}\boldsymbol{\theta}_S$ for some matrix of known coefficients $\boldsymbol{A}$. Let $\Omega_0 \subset \Omega_S$ denote the subspace of Ω_S where $\boldsymbol{\psi} = \mathbf{0}$. Let $\boldsymbol{\theta}_0$ denote an arbitrary elements of Ω_0. We wish to test the linear restriction $H_0 : \boldsymbol{\psi} = \mathbf{0}$. The Bayesian hypothesis test requires a prior distribution on Ω_S and a prior distribution on Ω_0. Let P denote the prior probability measure on Ω_S and let P_0 denote the prior probability measure on Ω_0. Let $\mathcal{M}_0$ denote the restricted model where $\boldsymbol{\psi} = \mathbf{0}$. The evidence under each model can be expressed as

$$p(\boldsymbol{y}|\mathcal{M}_0) = \int_{\Omega_0} L(\boldsymbol{\theta}_0)\ dP_0(\boldsymbol{\theta}_0)$$
$$p(\boldsymbol{y}|\mathcal{M}_S) = \int_{\Omega_S} L(\boldsymbol{\theta}_S)\ dP(\boldsymbol{\theta}_S).$$

We integrate with respect to the prior measures, as P_0 will not have density with respect to Lebesgue measure. The Bayes factor in favour of the null hypothesis, $\mathcal{B}_{01}$ is given by

$$\mathcal{B}_{01} = \frac{p(\boldsymbol{y}|\mathcal{M}_0)}{p(\boldsymbol{y}|\mathcal{M}_S)} \tag{2.17}$$
$$= \frac{\int_{\Omega_0} L(\boldsymbol{\theta}_0)\ dP_0(\boldsymbol{\theta}_0)}{\int_{\Omega_S} L(\boldsymbol{\theta}_S)\ dP(\boldsymbol{\theta}_S)}. \tag{2.18}$$

The Savage-Dickey density ratio occurs when we set the prior measure P_0 to be the conditional distribution of $\boldsymbol{\theta}_S$ given that $\boldsymbol{\psi} = \mathbf{0}$ under the encompassing model prior P. Let $P(\cdot|\boldsymbol{\psi} = \mathbf{0})$ be the regular conditional distribution on $\boldsymbol{\theta}_S$ given that $\boldsymbol{\psi} = \mathbf{0}$ under P. Regular conditional distributions are discussed in Billingsley (1999). Taking $P_0(\cdot) = P(\cdot|\boldsymbol{\psi} = \mathbf{0})$ gives that

$$p(\boldsymbol{y}|\mathcal{M}_0) = \int_{\Omega_0} L(\boldsymbol{\theta}_0)\ dP_0(\boldsymbol{\theta}_0)$$
$$= \int_{\Omega_0} L(\boldsymbol{\theta}_0)\ dP(\boldsymbol{\theta}_0|\boldsymbol{\psi} - \mathbf{0}). \tag{2.19}$$

Integrating with respect to the conditional distribution $P(\cdot|\boldsymbol{\psi} = \mathbf{0})$ also defines a conditional marginal likelihood under $\mathcal{M}_S$,

$$p(\boldsymbol{y}|\boldsymbol{\psi} = \mathbf{0}, \mathcal{M}_S) = \int_{\Omega_0} L(\boldsymbol{\theta}_0)\ dP(\boldsymbol{\theta}_0|\boldsymbol{\psi} = \mathbf{0}). \tag{2.20}$$

Substituting (2.20) and (2.19) into (2.17) gives another expression for the Bayes factor:

$$\log \mathcal{B}_{01} = \frac{p(\boldsymbol{y}|\boldsymbol{\psi} = \mathbf{0}, \mathcal{M}_S)}{p(\boldsymbol{y}|\mathcal{M}_S)}. \tag{2.21}$$

Dickey (1971) demonstrated that the the conditional evidence $p(\boldsymbol{y}|\boldsymbol{\psi} = \mathbf{0}, \mathcal{M}_S)$ can be expressed in terms of the posterior distribution of $\boldsymbol{\psi} = \boldsymbol{A}\boldsymbol{\theta}_S$ under the unconstrained model $\mathcal{M}_S$. The conditional evidence satisfies

$$p(\boldsymbol{y}|\boldsymbol{\psi} = \mathbf{0}, \mathcal{M}_S) = \frac{p(\boldsymbol{\psi} = \mathbf{0}|\boldsymbol{y}, \mathcal{M}_S)p(\boldsymbol{y}|\mathcal{M}_S)}{p(\boldsymbol{\psi} = \mathbf{0}|\mathcal{M}_S)}. \tag{2.22}$$

This does not follow immediately from Bayes' theorem the event $\boldsymbol{\psi} = \mathbf{0}$ is a subset of measure zero under P. The relationship (2.22) follows using the fact that $P(\cdot|\boldsymbol{\psi} = \mathbf{0})$ is a regular conditional distribution. Substituting (2.22) into (2.21) gives

$$\log \mathcal{B}_{01} = \frac{p(\boldsymbol{\psi} = \mathbf{0}|\boldsymbol{y}, \mathcal{M}_S)p(\boldsymbol{y}|\mathcal{M}_S)}{p(\boldsymbol{\psi} = \mathbf{0}|\mathcal{M}_S)} \frac{1}{p(\boldsymbol{y}|\mathcal{M}_S)} \tag{2.23}$$
$$= \frac{p(\boldsymbol{\psi} = \mathbf{0}|\boldsymbol{y}, \mathcal{M}_S)}{p(\boldsymbol{\psi} = \mathbf{0}|\mathcal{M}_S)}. \tag{2.24}$$

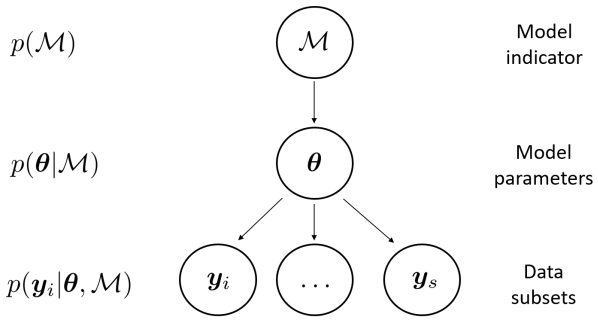


Figure 2.5: Target Bayesian model. Subsets are conditionally independent given the model $\mathcal{M}$ and the parameters $\boldsymbol{\theta}$, but conditionally dependent given only the model due to the shared dependence on $\boldsymbol{\theta}$.

The Bayes factor $\mathcal{B}_{01}$ is given by the ratio of the posterior density of $\boldsymbol{\psi}$ to the prior density of $\boldsymbol{\psi}$ under the unconstrained model $\mathcal{M}_S$. Evaluation of the density ratio is often easier than direct evaluation of the integrals in (2.18). A point of interest is that Savage-Dickey density ratio gives the integrated likelihood for the restricted model $\mathcal{M}_0$ in terms related to the unconstrained model $\mathcal{M}_S$,

$$p(\boldsymbol{y}|\mathcal{M}_0) = p(\boldsymbol{y}|\mathcal{M}_S)\frac{p(\boldsymbol{\psi} = \mathbf{0}|\boldsymbol{y}, \mathcal{M}_S)}{p(\boldsymbol{\psi} = \mathbf{0}|\mathcal{M}_S)}. \tag{2.25}$$

In many situations fitting the unconstrained model $\mathcal{M}_S$ will be considerably easier than fitting the constrained model $\mathcal{M}_0$. The identity (2.25) gives a useful alternative route for computing the integrated likelihood $p(\boldsymbol{y}|\mathcal{M}_0)$ in situations where it is difficult to fit the constrained model $\mathcal{M}_0$. In many situations we will be able to compute the evidence for the unconstrained model $\mathcal{M}_S$ and generate posterior samples from $\mathcal{M}_S$. Given posterior and prior samples of $\boldsymbol{\psi}$ we can use nonparametric kernel density estimation to estimation the density ratio in (2.24). This is assuming that the dimension of $\boldsymbol{\psi}$ is low, as the curse of dimensionality will rule out the feasibility of this approach in high-dimensions (Table 2.3). The evidence for the unconstrained model combined with the density ratio then give the evidence for the actual model of interest $\mathcal{M}_0$.

We will show the relationship in (2.25) has an important connection to the desired split-apply-combine methodology in Figure 2.1. We have a Big Data problem where it is difficult to fit the desired model on a single machine due to computational constraints so we adopt a divide and conquer strategy. In the next section we show how the split and apply steps effectively correspond to fitting an unconstrained model with an expanded parameter space. In the combine step we can compute a Bayes factor that connects the unconstrained model to the original model of interest. The analysis gives a Bayesian formalism to the relationship between the apply and combine steps in the embarrassingly parallel algorithm.

2.4 General Bayesian models

2.4.1 Introduction

Figure 2.5 diagrams the target model that we wish to use for Bayesian inference. When analysing the dataset using an embarrassingly parallel algorithm we obtain a collection of s subposterior distributions in the apply stage. We assume that worker returns the subposterior evidence $\widetilde{p}(\boldsymbol{y}_i)$ and samples from the subposterior $\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i)$ as output during apply stage for $i = 1, \ldots, s$. As demonstrated in the sleep dataset example in section 2.2.2 subposterior overlap is an important consideration in the combine step. Recall from (2.5) and (2.15) the following identity for the target model evidence,

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mathcal{M}) = \left(\prod_{i=1}^{s} \widetilde{p}(\boldsymbol{y}_i|\mathcal{M})\right) \alpha^s \int \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \mathcal{M})\ d\boldsymbol{\theta}. \tag{2.26}$$

$$= \left(\prod_{i=1}^{s} \widetilde{p}(\boldsymbol{y}_i|\mathcal{M})\right) \alpha^s \times I_{\text{sub}}. \tag{2.27}$$

To recover the target model evidence it is necessary evaluate the subposterior integral I_{sub}. In the case of $s = 2$ subsets, it is possible to see connection between the integral and a pointwise density measurement of subposterior overlap. To do so we rewrite the definition of the subposterior integral more explicitly as

$$I_{\text{sub}} = \int_{\boldsymbol{\theta}^* \in \Omega} \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta} = \boldsymbol{\theta}^* | \boldsymbol{y}_i, \mathcal{M}) \; d\boldsymbol{\theta}^*. \tag{2.28}$$

The extra notation is introduced as later it will be necessary to distinguish between the random variable in the target model $\boldsymbol{\theta}$ and an arbitrary ordinate in the parameter space $\boldsymbol{\theta}^*$. Each subposterior analysis returns a measure over the parameter space Ω. In (2.28) the variable of integration is now represented by the point $\boldsymbol{\theta}^*$ in the parameter space. Each subposterior density $\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_1), \ldots, \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_s)$ expresses different beliefs on the unknown parameter $\boldsymbol{\theta}$. Each subposterior thus assigns a different density to $\boldsymbol{\theta}^*$, and I_{sub} is calculated by taking the product over the s subsets and accumulating over the parameter space Ω.

With $s = 2$ subsets, we can see an immediate connection to the convolution of the two density functions $\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_1)$ and $\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_2)$. The convolution gives the density function of a random variable that can be defined as a difference of independent samples from each subposterior. Let $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$, be independent random variables with support Ω. Let the density function of $\boldsymbol{S}_1$ be given by the subposterior density from the first subset $\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_1)$ and the density function of $\boldsymbol{S}_2$ be given by the subposterior density from the second subset $\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_2)$. Now define the random variable $\boldsymbol{D} = \boldsymbol{S}_1 - \boldsymbol{S}_2$. The density of the difference random variable $\boldsymbol{D}$ at a point $\boldsymbol{r}$ is given by

$$f_D(\boldsymbol{r}) = \int_{\boldsymbol{\theta}^* \in \Omega} \widetilde{p}(\boldsymbol{\theta} = \boldsymbol{\theta}^* | \boldsymbol{y}_1, \mathcal{M}) \widetilde{p}(\boldsymbol{\theta} = \boldsymbol{\theta}^* - \boldsymbol{r} | \boldsymbol{y}_2, \mathcal{M}) \; d\boldsymbol{\theta}^*.$$

The density at $\mathbf{0}$ is equal to the subposterior integral I_{sub},

$$\begin{aligned} f_D(\mathbf{0}) &= \int_{\boldsymbol{\theta}^* \in \Omega} \widetilde{p}(\boldsymbol{\theta} = \boldsymbol{\theta}^* | \boldsymbol{y}_1, \mathcal{M}) \widetilde{p}(\boldsymbol{\theta} = \boldsymbol{\theta}^* | \boldsymbol{y}_2, \mathcal{M}) \; d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}^* \in \Omega} \prod_{i=1}^{2} \widetilde{p}(\boldsymbol{\theta} = \boldsymbol{\theta}^* | \boldsymbol{y}_i, \mathcal{M}) \; d\boldsymbol{\theta}^* \\ &= I_{\text{sub}}. \end{aligned} \tag{2.29}$$

If $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are distributed similarly, then we expect the difference $\boldsymbol{D}$ to be concentrated around zero. If the subposterior distributions are dissimilar then zero will be in the tails of the distribution of $\boldsymbol{D}$.

To illustrate, we consider another simple model choice problem. The dataset is from an experiment to detect extra sensory perception (ESP), first reported in Jahn et al. (1987). The experimenters constructed a random number generator using a radioactive source. The random number generator was calibrated to output a random sequence of zeroes and ones with equal probability. A subject was asked to psychically alter the sequence of random numbers using extra sensory capabilities. There are $n = 104,490,000$ observed random numbers which can be modelled as n independent Bernoulli(θ) events. This dataset is also considered in Bernardo et al. (2011). Table 2.4 reports the number of ones in the observed sequence (successes). The fraction of ones in the dataset is slightly higher than 0.5.

We compare two models, $\mathcal{M}_1 : \theta = 0.5$, corresponding to no support for ESP and $\mathcal{M}_2 : \theta \neq 0.5$, which allows for the possibility for ESP. We use a flat Beta$(1, 1)$ prior on the proportion θ for model 2. The ESP dataset is commonly used to illustrate differences between Bayesian and Frequentist hypothesis testing. Suppose we assign equal prior weight to each model. The posterior probability of $\mathcal{M}_1 : \theta = 0.5$ is then 0.92. The p-value for testing $\theta = 0.5$ is 0.0003017. The Bayesian approach favours the simpler model $\mathcal{M}_1 : \theta = 0.5$, when the frequentist analysis rejects the null hypothesis ($\theta = 0.5$) at the usual 5 percent level of significance. Here default Bayesian and frequentist procedures lead to opposite conclusions regarding the existence of extra sensory perception. We use the ESP dataset to illustrate divide and conquer inference as it is a simple large n dataset where the posterior model probabilities are not highly concentrated around zero or one.

	Full dataset	Subset 1	Subset 2
Trials	104,490,000	52,245,000	52,245,000
Successes	52,263,471	26,137,735	26,125,736
Success proportion	0.5001768	0.5002916	0.5000619

Table 2.4: Full ESP dataset and subsets for a divide and conquer approach with $s = 2$. The success proportion is close to 0.5 in the full dataset and each subset.

We split the full dataset into two subsets, as described in Table 2.4. Let $\boldsymbol{y}_1$ denote the data for subset 1 and $\boldsymbol{y}_2$ denote the data in subset 2. We focus on the divide and conquer analysis of model 2, where we estimate the success probability. Let n_i and r_i denote the number of trials and successes respectively for subset $i = 1, 2$. The subpriors for θ are both uniform Beta$(1, 1)$ since a fractionated uniform density is proportional to a uniform density: $[\text{Beta}(1, 1)(\theta)]^{1/2} = \mathbf{1}_{\theta\in[0,1]} = \text{Beta}(1, 1)(\theta)$. Given a Beta$(1, 1)$ subpriors, each subposterior is a Beta$(r_i + 1, n_i - r_i + 1)$ density for $i = 1, 2$. Figure 2.6 plots the two subposteriors. As each subposterior is a Beta density it is possible to work out the subposterior integral analytically. Let $a_i = r_i + 1, b_i = n_i - r_i + 1, a^* = (\sum_{i=1}^s a_i) - (s - 1)$ and $b^* = (\sum_{i=1}^s b_i) - (s - 1)$. The subposterior integral satisfies

$$\log I_{\text{sub}} = \log B(a^*, b^*) - \sum_{i=1}^{s} \log B(a_i, b_i), \tag{2.30}$$

where $B(a, b)$ is the Beta function. Using the data partition in Table 2.4, we calculate $I_{\text{sub}} = 259.64$. Panel (b) shows the density of $D = S_1 - S_2$, where S_1 and S_2 are distributed according to $\widetilde{p}(\theta|\boldsymbol{y}_1)$ and $\widetilde{p}(\theta|\boldsymbol{y}_2)$ respectively. The value of the subposterior integral $I_{\text{sub}} = 259.64$ is plotted as a dashed horizontal line and the vertical dotted line has an x intercept of zero. Looking at the intersection we see that the theoretical density of D at zero is equal to I_{sub} giving an empirical verification of (2.29) for this example. The convolution by inspection argument leading to (2.29) was non-constructive, and it is not clear how to generalise the argument to an arbitrary number of subsets s. Subposterior overlap clearly has a role in the combine step, although it is not apparent if this is an inherently Bayesian approach to the evidence task in the final stage. Another open question is the significance of the subprior normalising constant α in the representation (2.5). We take a more general approach in the next subsection that addresses these issues.

2.4.2 Subset saturated model

In the divide and conquer analysis, each data subset $\boldsymbol{y}_i$ is modelled separately using the subprior $\widetilde{p}(\boldsymbol{\theta})$. This Big Data modelling strategy is adopted for purely computational reasons. This strategy can also be motivated through an inferential argument. Laying out the inferential argument leads to a Savage-Dickey ratio representation in the combine step. Consider an alternative hierarchical model, diagrammed in Figure 2.7, where each data subset is now allocated an independent set of parameters. As before, suppose that the arbitrary model $\mathcal{M}$ is indexed by parameter $\boldsymbol{\theta}$, where $\boldsymbol{\theta} \in \Omega$. For each model $\mathcal{M}$ we introduce an expanded version $\mathcal{M}_S$ that has a separate set of parameters for each subset. In the alternative hierarchical model we have an expanded parameter space. Let $\Omega_S = \prod_{i=1}^s \Omega$. Let $\boldsymbol{\theta}_S \in \Omega_S$ denote the parameter for the expanded model. Then

$$\boldsymbol{\theta}_S = \begin{bmatrix} \boldsymbol{\theta}^{(1)} \\ \vdots \\ \boldsymbol{\theta}^{(s)} \end{bmatrix} \in \Omega_S.$$

Suppose the original model $\mathcal{M}$ is a regression model with a particular set of explanatory variables. The expanded model $\mathcal{M}_S$ will use the same set of covariates in each subset, but allows for different coefficients

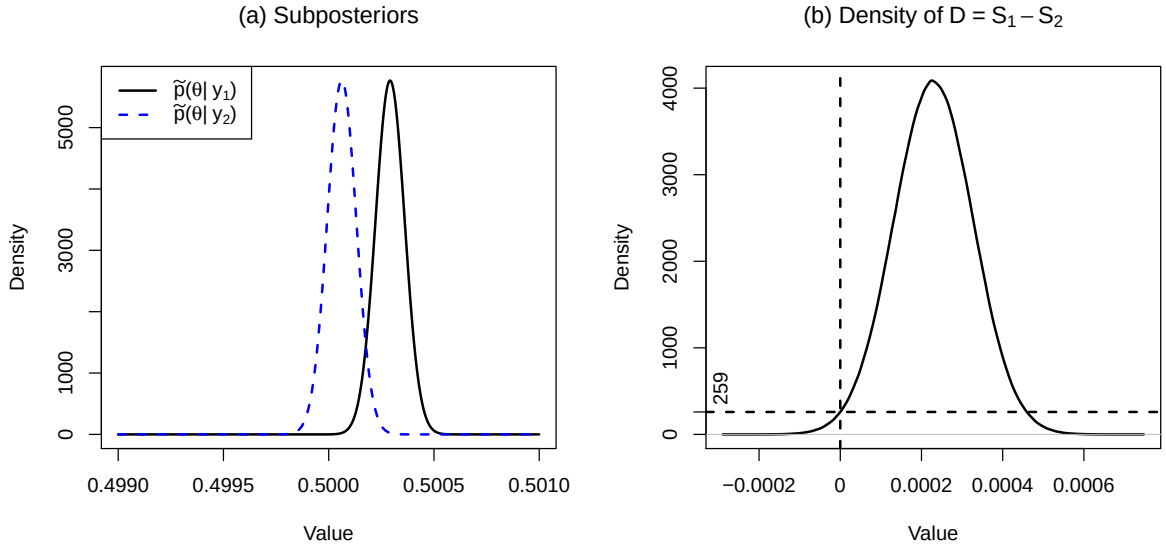


Figure 2.6: Illustration of the subposterior density identity using the ESP dataset ($s = 2$). The full dataset and data subsets are summarised in Table 2.4. (a) Subposterior densities. (b) Density of the difference variable D. In (b) the horizontal dashed line gives the value of I_{sub}. The intersection of the dashed lines in (b) illustrates the density identity (2.29).

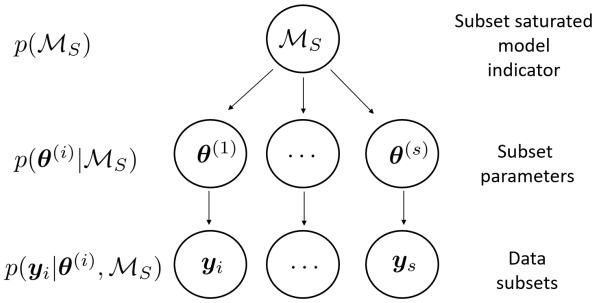


Figure 2.7: Alternative hierarchical Bayesian model (subset saturated model). Each subset $\boldsymbol{y}_i$ is allocated an independent set of parameters $\boldsymbol{\theta}^{(i)}$. Subsets are conditionally independent gives the model indicator. The split and apply steps in the divide and conquer procedure are equivalent to defining and fitting the subset saturated model.

in each subset. It can be argued that split and apply steps in the embarrassingly parallel algorithm are effectively the same as fitting the alternative hierarchical model with subset specific parameters. The alternative hierarchical model is conceptually similar to the saturated model that is used when calculating the deviance in a generalised linear model. The saturated model in the context of generalised linear models allows for a unique parameter for each observation and provides the best possible goodness of fit. In the divide and conquer world, the expanded model $\mathcal{M}_S$ allows for a unique parameter for each subset, and allows extra adaptation compared to the desired model $\mathcal{M}$. As such we will refer to $\mathcal{M}_S$ as the subset saturated model, and the hierarchical model in Figure 2.7 as the hierarchical subset saturated model.

2.4.3 Split and apply steps

We make some assumptions about the subset saturated model to link it concretely to the split and apply steps that take place in the embarrassingly parallel algorithm (see Figure 2.1).

Assumption 1: Subset likelihoods are the same as in the target model. That is for all $\boldsymbol{\theta}^* \in \Omega$, $p(\boldsymbol{y}_i|\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^*, \mathcal{M}_S) = p(\boldsymbol{y}_i|\boldsymbol{\theta} = \boldsymbol{\theta}^*, \mathcal{M})$ for $i = 1, \ldots, s$.

Assumption 2: The subset specific parameters have independent priors, so the joint prior in the subset

saturated model can be written as

$$p(\boldsymbol{\theta}^{(i)}, \ldots, \boldsymbol{\theta}^{(s)}|\mathcal{M}_S) = \prod_{i=1}^{s} p(\boldsymbol{\theta}^{(i)}|\mathcal{M}_S).$$

Assumption 3: Each subset prior is given by the subprior used in a divide and conquer analysis (see Table 2.1). That is for all $\boldsymbol{\theta}^* \in \Omega$, $p(\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^*|\mathcal{M}_S)$ is equal to $\widetilde{p}(\boldsymbol{\theta} = \boldsymbol{\theta}^*|\mathcal{M})$ for $i = 1, \ldots, s$.

The subset saturated model is a bona fide Bayesian model for any choice of prior distribution on the subset specific parameters $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(s)}$. Assumptions 2 and 3 are needed to make a connection to the divide and conquer procedure. Assumption 1 is needed to be completely specify the model. Let $L_S^{(i)}(\boldsymbol{\theta})$ give the likelihood function for subset i under the subset saturated model (Figure 2.7). Let $L^{(i)}(\boldsymbol{\theta})$ give the likelihood function for subset i under the target model $\mathcal{M}$ (Figure 2.5). From Assumption 1 $L_S^{(i)}(\boldsymbol{\theta}) = L^{(i)}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Omega$.

Under assumptions 1,2 and 3, the posterior distribution of the subset specific parameters $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(s)}$ in the subset saturated model is closely related to the subposterior output that is is generated in the apply phase. The underlying space of interest is Ω. The apply stage in the embarrassingly parallel algorithm returns s belief distributions over Ω. The subset saturated model also defines s belief distributions on Ω through the s marginal distributions on the subset specific parameter $\boldsymbol{\theta}^{(i)}$ for $i = 1, \ldots, s$. We can show a measure theoretic equivalence between the distribution sets of the embarrassingly parallel algorithm and the marginal distributions of the subset saturated model under assumptions 1,2 and 3. Let $\mathcal{F}$ denote the Borel σ-algebra on Ω. Each worker in the apply step is given an initial belief distribution on Ω given by the subprior density $\widetilde{p}(\boldsymbol{\theta}|\mathcal{M})$. Let $\widetilde{P}^{(i)}$ denote the subprior distribution for worker i for $i = 1, \ldots, s$. In the subset saturated model we have a collection of s prior distributions on Ω, one for each subset specific parameter. By assumption 3 each subset specific prior is set to the subprior distribution. Specifically, we have that for all sets $F \in \mathcal{F}$:

$$\widetilde{P}^{(i)}(F) = P^{(i)}(F) \quad i = 1, \ldots, s. \tag{2.31}$$

In the embarrassingly parallel algorithm each worker learns updated beliefs on Ω during the apply step. Let $\widetilde{P}_\pi^{(i)}$ denote the ith subposterior distribution on Ω. The subset saturated model has parameter space $\Omega_S = \prod_{i=1}^{s} \Omega$. Let P denote the prior distribution on Ω_S under the subset saturated model and let P_π denote the posterior distribution on Ω_s under the subset saturated model. Let $P_\pi^{(i)}$ be the marginal posterior distribution on $\boldsymbol{\theta}^{(i)}$ under the subset saturated model for $i = 1, \ldots, s$. Using the subset saturated model we also obtain a collection of beliefs on Ω through the s marginal distributions $P_\pi^{(i)}$ for $i = 1, \ldots, s$. From Assumption 1 and 3 we have that

$$\widetilde{P}_\pi^{(i)}(F) = P_\pi^{(i)}(F) \quad i = 1, \ldots, s. \tag{2.32}$$

Let μ denote Lebesgue measure on $\mathbb{R}^d$ where d is the dimension of Ω. Under Assumptions 1 and 3 both measures have the same Radon-Nikodym derivative with respect to Lebesgue measure,

$$\frac{\widetilde{P}_\pi^{(i)}}{d\mu} = L_S^{(i)}\widetilde{p}(\boldsymbol{\theta}) = L^{(i)}(\boldsymbol{\theta})\widetilde{p}(\boldsymbol{\theta})$$
$$\frac{P_\pi^{(i)}}{d\mu} = L^{(i)}(\boldsymbol{\theta})\widetilde{p}(\boldsymbol{\theta}).$$

This is sufficient to establish (2.32). The split and apply steps in the embarrassingly parallel algorithm (see Figure 2.1) are effectively the same as building and computing the posterior distribution for the subset saturated model (Figure 2.7). The subset saturated model is also useful for describing the role of the the subposterior integral (2.15) in the combine step.

2.4.4 Combine step

The ideal Bayesian procedure in the combine step can be identified by considering a linearly restricted version of the subset saturated model (Figure 2.7). The key observation is that the target model $\mathcal{M}$ is effectively nested within the subset saturated model $\mathcal{M}_S$ under the linear restriction that $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(2)} = \ldots = \boldsymbol{\theta}^{(s)}$. Let $\boldsymbol{\psi}$ denote a $d(s-1)$ dimensional vector that captures differences in parameter values across subsets, specifically

$$\boldsymbol{\psi} = \begin{bmatrix} \boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(2)} \\ \boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(3)} \\ \vdots \\ \boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(s)} \end{bmatrix}. \tag{2.33}$$

Under the linear restriction $\boldsymbol{\psi} = \mathbf{0}$ the structure of the subset saturated model diagrammed in Figure 2.7 collapses to that of the target hierarchical model diagrammed in Figure 2.5. We will argue that the model evidence for the linearly restricted version of the subset saturated model is equal to the model evidence for the target model under appropriate assumptions.

There are some measure theoretic considerations when working with the linearly restricted version of the subset saturated model. Again let P denote the prior measure on $\Omega_S = \prod_{i=1}^{s} \Omega$ given that assumptions 1,2 and 3 hold. Let Ω_0 be the linear subspace of Ω_S defined by the linear restriction $\boldsymbol{\psi} = \mathbf{0}$. Let $P(\cdot|\boldsymbol{\psi} = \mathbf{0})$ denote the regular conditional distribution of $\boldsymbol{\theta}_S$ given that $\boldsymbol{\psi} = \mathbf{0}$ under P. Let $\mathcal{M}_0$ denote the linearly restricted model where $\boldsymbol{\theta}_S \in \Omega_0$. Let $\boldsymbol{\theta}_0$ denote an element of $\boldsymbol{\Omega}_0$. The linearly restricted model $\mathcal{M}_0$ is not completely equivalent to the target model $\mathcal{M}$ as the parameter spaces are different. Model $\mathcal{M}_0$ puts a prior distribution P_0 on $\Omega_0 \subset \prod_{i=1}^{s} \Omega$ whereas the target model $\mathcal{M}$ puts a prior distribution on Ω.

Suppose we set $P_0 \equiv P(\cdot|\boldsymbol{\psi} = \mathbf{0})$. Let $P_0^{(1)}$ be the marginal distribution on $\boldsymbol{\theta}^{(1)}$ under P_0. Let $P_{\mathcal{M}}$ be the prior distribution on Ω under the target model. The measure $P_{\mathcal{M}}$ has density $p(\boldsymbol{\theta}|\mathcal{M})$ with respect to Lebesgue measure on $\mathbb{R}^d$. Assumptions 1,2 and 3 have been selected so that $P_0^{(1)}$ and $P_{\mathcal{M}}$ give the same distribution on Ω. More formally, we can show that

$$P_0^{(1)}(F) = P_{\mathcal{M}}(F), \tag{2.34}$$

for all sets $F \in \mathcal{F}$ where $\mathcal{F}$ is the Borel σ-algebra on Ω. Let $L_0(\boldsymbol{\theta}_0)$ denote the likelihood function for the linearly restricted model $\mathcal{M}_0$ and let $L(\boldsymbol{\theta})$ denote the likelihood function for the target model $\mathcal{M}$. Suppose that we have some $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(s)}) \in \Omega_0$. On Ω_0 all subset specific parameters are equal. If $\boldsymbol{\theta}_0 \in \Omega_0$ then $L_0(\boldsymbol{\theta}_0) = L(\boldsymbol{\theta}^{(1)})$. This follows from Assumption 1, the subset specific likelihoods are the same as the target model likelihood for each subset.The model evidence for $\mathcal{M}_0$ is directly expressed as an integral over Ω_0. We can also express the model evidence for $\mathcal{M}_0$ as an integral over Ω by considering the marginal distribution of $\boldsymbol{\theta}^{(1)}$ under P_0.

$$\begin{aligned} p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mathcal{M}_0) &= \int_{\boldsymbol{\theta}_0 \in \Omega_0} L_0(\boldsymbol{\theta}_0)\ dP_0(\boldsymbol{\theta}_0) \\ &= \int_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta})\ dP_0^{(1)}(\boldsymbol{\theta}). \end{aligned} \tag{2.35}$$

The evidence for the linearly restricted model $\mathcal{M}_0$ can also be obtained by integrating the target model likelihood function $L(\boldsymbol{\theta})$ with respect to $P_0^{(1)}(\boldsymbol{\theta})$. Under assumptions 1,2 and 3 we have the equality (2.34). We can therefore substitute the target model prior distribution $P_{\mathcal{M}}(F)$ into (2.35). The target model prior distribution $P_{\mathcal{M}}(\cdot)$ has density $p(\boldsymbol{\theta}|\mathcal{M})$ with respect to Lebesgue measure. The evidence for

the linearly restricted model is then shown to be equal to the evidence for the target model:

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mathcal{M}_0) = \int_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta})\ dP_{\mathcal{M}}(\boldsymbol{\theta}) \tag{2.36}$$

$$= \int_{\boldsymbol{\theta}\in\Omega} p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_S|\mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M})\ d\boldsymbol{\theta} \tag{2.37}$$

$$= p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mathcal{M}). \tag{2.38}$$

Using the Savage-Dickey density ratio (2.22), we can express the model evidence for the linearly restricted model $\mathcal{M}_0$ in terms of the subset saturated model $\mathcal{M}_S$:

$$\begin{aligned} p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mathcal{M}_0) &= p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\boldsymbol{\psi} = \mathbf{0}, \mathcal{M}_S) \\ &= p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mathcal{M}_S) \frac{p(\boldsymbol{\psi} = \mathbf{0}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s, \mathcal{M}_S)}{p(\boldsymbol{\psi} = \mathbf{0}|\mathcal{M}_S)}. \end{aligned} \tag{2.39}$$

Now using the fact that $p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mathcal{M}_0) = p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mathcal{M})$, we have the following identity relating the target model evidence to the subset saturated model:

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mathcal{M}) = p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mathcal{M}_S) \frac{p(\boldsymbol{\psi} = \mathbf{0}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s, \mathcal{M}_S)}{p(\boldsymbol{\psi} = \mathbf{0}|\mathcal{M}_S)}. \tag{2.40}$$

We will show that each of the terms on the right hand side of (2.40) can be estimated using a split-apply-combine algorithm. We assume that each worker can compute the subposterior evidence score $\widetilde{p}(\boldsymbol{y}_i|\mathcal{M})$ for $i = 1, \ldots, s$. The local evidence on each batch of data can be computed using existing techniques. Existing methods for evidence estimation are discussed in Chapter 3. We now show how the density ratio can be estimated in the combine stage using subposterior samples generated in the apply stage. The posterior distribution $p(\boldsymbol{\psi}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s, \mathcal{M}_S)$ is effectively given by the subposterior distributions in a divide and conquer analysis. The prior distribution $p(\boldsymbol{\psi}|\mathcal{M}_S)$ is effectively given by the subprior distribution $\widetilde{p}(\boldsymbol{\theta})$ (see Table 2.1). Define the random variables $\boldsymbol{S}_i \sim \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \mathcal{M})$ and $\boldsymbol{V}_i \sim \widetilde{p}(\boldsymbol{\theta}, \mathcal{M})$ for $i = 1, \ldots, s$. The random variable $\boldsymbol{S}_i$ is distributed according to the ith subposterior, and the random variable $\boldsymbol{V}_i$ is distributed according to the ith subprior.

Now define the random vectors $\boldsymbol{D}$ and $\boldsymbol{D}_0$ as

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{S}_1 - \boldsymbol{S}_2 \\ \ldots \\ \boldsymbol{S}_1 - \boldsymbol{S}_s \end{bmatrix}, \quad \boldsymbol{D}_0 = \begin{bmatrix} \boldsymbol{V}_1 - \boldsymbol{V}_2 \\ \ldots \\ \boldsymbol{V}_1 - \boldsymbol{V}_s \end{bmatrix}. \tag{2.41}$$

Then $\boldsymbol{D} \sim p(\boldsymbol{\psi}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s, \mathcal{M}_S)$ and $\boldsymbol{D}_0 \sim p(\boldsymbol{\psi}|\mathcal{M}_S)$. We can sample from the subprior and subposterior distributions in the apply step. Using the definitions in (2.41) we can sample from $p(\boldsymbol{\psi}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s, \mathcal{M}_S)$ and $p(\boldsymbol{\psi}|\mathcal{M}_S)$ in the combine step by pooling the subposterior and subprior samples. The remaining term on the right hand side of (2.39) is the evidence for the subset saturated model $p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mathcal{M}_S)$. This can easily be computed given the subposterior evidence values $\widetilde{p}(\boldsymbol{y}_1|\mathcal{M}), \ldots, \widetilde{p}(\boldsymbol{y}_s|\mathcal{M})$.

The model evidence for the subset saturated model is the product of the subposterior evidence scores.

Specifically,

$$\begin{aligned} p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mathcal{M}_S) &= \int_{\Omega_s} L(\boldsymbol{\theta}_S)\ dP(\boldsymbol{\theta}_S) \\ &= \int_{\Omega_s} \left(\prod_{i=1}^{s} L_S^{(i)}(\boldsymbol{\theta}^{(i)}) \right) \prod_{i=1}^{s} dP^{(i)}(\boldsymbol{\theta}^{(i)}) \\ &= \prod_{i=1}^{s} \int_{\Omega} L_S^{(i)}(\boldsymbol{\theta})\ dP^{(i)}(\boldsymbol{\theta}) \\ &= \prod_{i=1}^{s} \int_{\Omega} L^{(i)}(\boldsymbol{\theta})\ d\widetilde{P}(\boldsymbol{\theta}) \\ &= \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{y}_i|\mathcal{M}). \end{aligned}$$

Substituting the previous result into (2.39), we obtain the following identity for the full dataset integrated likelihood:

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mathcal{M}) = \left(\prod_{i=1}^{s} \widetilde{p}(\boldsymbol{y}_i|\mathcal{M}) \right) \frac{p(\boldsymbol{\psi} = \mathbf{0}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s, \mathcal{M}_S)}{p(\boldsymbol{\psi} = \mathbf{0}|\mathcal{M}_S)}. \tag{2.42}$$

The density ratio in (2.42) can be interpreted as the posterior to prior odds of observing $\boldsymbol{\psi} = \mathbf{0}$ in the subset saturated model. If the model is appropriate for the entire dataset, we expect to see similar subposterior distributions on $\boldsymbol{\theta}$ over subsets, and the posterior density of $\boldsymbol{\psi}$ should be concentrated around zero. What exactly constitutes a 'high' level of agreement is determined by comparing to the prior density of $\boldsymbol{\psi}$ at zero. The ratio of the two densities has an interpretation as a Bayes factor in the subset saturated model. Recall the identity in the introduction,

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mathcal{M}) = \left(\prod_{i=1}^{s} \widetilde{p}(\boldsymbol{y}_i|\mathcal{M}) \right) \alpha^s \int \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \mathcal{M})\ d\boldsymbol{\theta} \tag{2.43}$$

Comparing (2.42) to (2.43) we see a correspondence between the density ratio and the subposterior integral and subprior normalising constant. Specifically, it can be shown that

$$p(\boldsymbol{\psi} = \mathbf{0}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s, \mathcal{M}_S) = \int \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i)\ d\boldsymbol{\theta} = I_{\text{sub}} \tag{2.44}$$

$$p(\boldsymbol{\psi} = \mathbf{0}|\mathcal{M}_S) = \alpha^s. \tag{2.45}$$

Equation (2.44) is the generalisation of the density identity for $s = 2$ presented given in (2.29). Equation (2.45) helps to explain why the subprior normalising constant α appears in the full model evidence decomposition (2.43). The ratio of I_{sub} to α^s measures the consistency of parameter estimates across subsets. The split-apply-combine approach for computing the evidence can be viewed as defining a misspecified model in the split step, followed by fitting the misspecified model in the apply step. The misspecified model can be viewed as a subset saturated model where we have introduced extra parameters. The combine step then corrects for the misspecification by computing an adjustment that corresponds to a Bayes factor linking the subset saturated model to the target model.

2.4.5 Example: ESP dataset

To illustrate the results in the previous subsection, we present another analysis of the ESP dataset. We split the dataset into three subsets, as described in Table 2.5. Let $\boldsymbol{y}_i$ denote the data in subset i for $i = 1, 2, 3$. Additionally, let n_i denotes the number of trials in subset i and r_i denote the number of successes for $i = 1, 2, 3$.

	Full dataset	Subset 1	Subset 2	Subset 3
Trials	104,490,000	34,830,000	34,830,000	34,830,000
Successes	52,263,471	17,421,157	17,415,157	17,427,157
Success proportion	0.5001768	0.5001768	0.5000045	0.5003490

Table 2.5: Full ESP dataset and subsets for a divide and conquer approach with $s = 3$. The success proportion is close to 0.5 in the full dataset and in each subset.

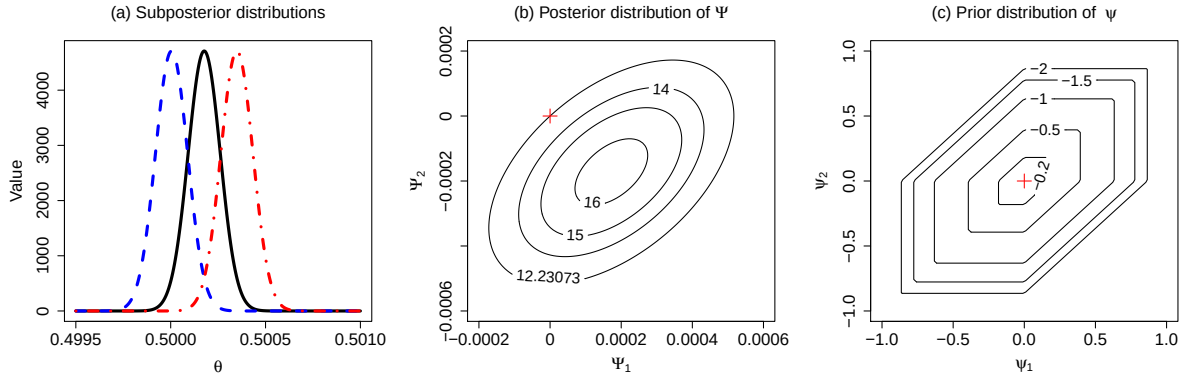


Figure 2.8: (a) Subposterior distributions for $s = 3$ subsets. (b) Posterior distribution of ψ. (c) Prior distribution of ψ. The red cross in (b) and (c) denotes the point (0,0). The numeric labels on the contours in (b) and (c) give the log density on the respective contour. Subposterior overlap is measured by the posterior density of ψ at zero. This is compared to the prior density of ψ at zero to determine the evidence for a linearly restricted model. The positioning of the red cross in panels (b) and (c) illustrate the identities (2.44) and (2.45) respectively.

We again use a flat prior on the unknown success probability. The subprior is a Beta$(1, 1)$ density. Each subposterior is a Beta$(1 + r_i, 1 + n_i - r_i)$ distribution for $i = 1, 2, 3$. Using (2.30) we can calculate

$$\log I_{\text{sub}} = 12.2307$$
$$\log \alpha = 0.$$

Given that $s = 3$, $\boldsymbol{\psi}$ is a two dimensional vector,

$$\boldsymbol{\psi} = \begin{bmatrix} \theta^{(1)} - \theta^{(2)} \\ \theta^{(1)} - \theta^{(3)} \end{bmatrix},$$

where $\theta^{(i)}$ is the proportion parameter in subset i for $i = 1, 2, 3$. Figure 2.8 illustrates the identities in (2.44) and (2.45). Panel (a) shows the three subposterior distributions. Panel (b) displays contours of the posterior distribution of $\boldsymbol{\psi}$. The red cross denotes the point $(0, 0)$. The density contour where $\log p(\boldsymbol{\psi}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s) = 12.23$ passes through the point $(0, 0)$ illustrating the general result that $p(\boldsymbol{\psi} = \mathbf{0}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s, \mathcal{M}) = \int \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \mathcal{M})\, d\boldsymbol{\theta}$. Panel (c) shows the contours of the prior density of $\boldsymbol{\psi}$. The prior mode is at $(0, 0)$, and it can be seen that the log density is approaching zero around this point. We thus have empirical evaluation of (2.44) and (2.45) in this example.

As mentioned, it is possible to generate samples from $p(\boldsymbol{\psi}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s, \mathcal{M}_S)$. Evaluation of the subposterior integral can then be approached as a pointwise density estimation task $\widehat{I}_{\text{sub}} = \widehat{p}(\boldsymbol{\psi} = \mathbf{0}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s, \mathcal{M}_S)$. From this example we can identify an important issue regarding the impact of the partition made in the *split* step. In (a) we can see that there is only a moderate amount of subposterior overlap. As such, $(0, 0)$ is in the tails of the posterior distribution of $\boldsymbol{\psi}$. The red cross in panel (b) is in the tails of the posterior distribution on $\boldsymbol{\psi}$. Estimating the pointwise density at zero will be difficult if $\mathbf{0}$ is in a region with little posterior support. This promotes splitting the data so that subposterior distributions are as similar as possible. Although judicious data partitioning can shift the posterior support of $\boldsymbol{\psi}$ to a more favourable

region, little can be done about the fact that $\boldsymbol{\psi}$ is a $d(s-1)$-dimensional vector. The implications for estimation of I_{sub} are discussed in the next subsection.

2.4.6 Curse of dimensionality

Suppose that we have B samples from each subposterior. Let $\boldsymbol{S}_i^{[b]}$ be the bth sample from $\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \mathcal{M})$ for $b = 1, \ldots, B$ and $i = 1, \ldots, s$. We can generate B posterior samples of $\boldsymbol{\psi}$, by setting

$$\boldsymbol{\psi}^{[b]} = \begin{bmatrix} \boldsymbol{S}_1^{[b]} - \boldsymbol{S}_2^{[b]} \\ \boldsymbol{S}_1^{[b]} - \boldsymbol{S}_3^{[b]} \\ \vdots \\ \boldsymbol{S}_1^{[b]} - \boldsymbol{S}_s^{[b]} \end{bmatrix}, \tag{2.46}$$

for $b = 1, \ldots, B$. Using any consistent kernel density estimator the subposterior integral can be estimated as

$$\widehat{I}_{\text{sub}} = \widehat{p}(\boldsymbol{\psi} = \mathbf{0}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s, \mathcal{M}_S) = \frac{1}{B}\sum_{b=1}^{B} K_{\boldsymbol{H}}(\boldsymbol{\psi}^{[b]}), \tag{2.47}$$

for a suitable kernel density estimator $K_{\boldsymbol{H}}(\cdot)$ with bandwidth matrix $\boldsymbol{H}$.

Although the estimator (2.47) is simulation consistent, it will be heavily affected by the curse of dimensionality. As mentioned in Section 2.3.1, the optimal mean square error rate for nonparametric kernel density estimation is known to be $O(B^{-2/(q+4)})$, where q is the dimension of the space (Terrell and Scott, 1992). As the number of subsets s increases, the dimension of $\boldsymbol{\psi}$ increases linearly. Non parametric kernel density estimation in a $(s-1)d$ dimensional space will only be feasible for very small s and d (recall Table 2.3). This suggests that the Monte Carlo budget per worker needs to increase exponentially with the number of subsets s to control the error in the combine step.

2.4.7 Embarrassingly parallel evidence estimation

Algorithm 2.1 gives the general algorithm for computing the model evidence in parallel using the kernel density estimator for the subposterior integral (2.47). We expect the algorithm to break down as s increases, as the error in the combine stage will increase exponentially with s for a fixed Monte Carlo budget B. This limits its use for scalable computation of the integrated likelihood. The computational benefits of increasing s in the initial split are undermined by the increased Monte Carlo error in the combine step. We thus turn to a different strategy involving data augmentation.

2.5 Data augmentation for distributed inference

2.5.1 Introduction

The general treatment of the problem given in the previous section is interesting, but leads to a somewhat negative conclusion regarding the feasibility of the divide and conquer approach. We can reach a more positive outcome by working within the confines of the conditionally conjugate exponential family. To ease notation we do not explicitly condition on a particular model $\mathcal{M}$ in this section. The overall strategy makes use of Gibbs sampling in the apply step so that we can implement an efficient estimator of the subposterior integral in the combine step. The data augmentation based algorithm has more favourable scaling properties than the kernel density based algorithm (Algorithm 2.1). There is a connection between our approach and Chib's method (Chib, 1995) for estimating the marginal likelihood from the Gibbs output in the single machine setting.

Algorithm 2.1 Divide and Conquer model evidence

Apply Step: Subset Markov chain Monte Carlo (MCMC) runs
for $i = 1, \ldots, s$ **do**
  for $b = 1, \ldots, B$ **do**
    Sample $\boldsymbol{\theta}$ from $\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i)$
    Set $\boldsymbol{S}_i^{[b]} \leftarrow \boldsymbol{\theta}$
  end for
  Compute $\widehat{\log}\ \widetilde{p}(\boldsymbol{y}_i)$ using method of choice.
end for

Combine Step: Post Processing subset MCMC output
for $b = 1, \ldots, B$ **do**
  Compute $\boldsymbol{\psi}^{[b]} = \begin{bmatrix} \boldsymbol{S}_1^{[b]} - \boldsymbol{S}_2^{[b]} \\ \boldsymbol{S}_1^{[b]} - \boldsymbol{S}_3^{[b]} \\ \vdots \\ \boldsymbol{S}_1^{[b]} - \boldsymbol{S}_s^{[b]} \end{bmatrix}$.
end for
Compute $\widehat{I}_{\text{sub}} = B^{-1} \sum_{b=1}^{B} K_{\boldsymbol{H}}(\boldsymbol{\psi}^{[b]})$
Compute $\log \widehat{p}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s) = \sum_{i=1}^{s} \widehat{\log}\ \widetilde{p}(\boldsymbol{y}_i) + s \log \alpha + \widehat{I}_{\text{sub}}$

2.5.2 Conjugate priors in the exponential family

To describe the conjugate priors suppose we have n independently and identically distributed observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$. Bold face is used to allow for vector values observations. The model conditioned likelihood belongs to the exponential family if the likelihood contribution for a single observation can be written as

$$p(\boldsymbol{y}_i|\boldsymbol{\theta}) = h(\boldsymbol{y}_i) g(\boldsymbol{\theta}) \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} t(\boldsymbol{y}_i)), \tag{2.48}$$

for some known functions $h(\cdot), g(\cdot), t(\cdot)$ and $\boldsymbol{\eta}(\cdot)$. The function $t(\cdot)$ returns a vector of sufficient statistics and the function $\boldsymbol{\eta}(\cdot)$ dictates how the parameter $\boldsymbol{\theta}$ interacts with the data. The standard conjugate prior (Bernardo and Smith, 2006) on $\boldsymbol{\theta}$ is parametrised by a scalar ν_0, and a vector $\boldsymbol{\phi}_0$ of the same dimension as $t(\boldsymbol{y}_i)$, taking the form

$$\pi(\boldsymbol{\theta}; \nu_0, \boldsymbol{\phi}_0) = c(\nu_0, \boldsymbol{\phi}_0) g(\boldsymbol{\theta})^{\nu_0} \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{\phi}_0). \tag{2.49}$$

The function $g(\cdot)$ is the same that appears in the data likelihood (2.48), and $c(\nu_0, \boldsymbol{\phi}_0)$ is an appropriate normalising constant:

$$\frac{1}{c(\nu_0, \boldsymbol{\phi}_0)} = \int g(\boldsymbol{\theta})^{\nu_0} \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{\phi}_0)\ d\boldsymbol{\theta}.$$

Explicit forms of some standard conjugate priors are listed in Chapter 5 of Bernardo and Smith (2006). Suppose the prior is of the standard conjugate form, so $p(\boldsymbol{\theta}) = c(\nu_0, \boldsymbol{\phi}_0) g(\boldsymbol{\theta})^{\nu_0} \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{\phi}_0)$. Given n independent and identically distributed observations from model (2.48), the posterior remains in the same family,

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) &\propto \left(\prod_{i=1}^{n} h(\boldsymbol{y}_i) g(\boldsymbol{\theta}) \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} t(\boldsymbol{y}_i)) \right) c(\nu_0, \boldsymbol{\phi}_0) g(\boldsymbol{\theta})^{\nu_0} \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{\phi}_0) \\ &\propto g(\boldsymbol{\theta})^{n} \exp\left[\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} (\textstyle\sum_{i=1}^{n} t(\boldsymbol{y}_i)) \right] c(\nu_0, \boldsymbol{\phi}_0) g(\boldsymbol{\theta})^{\nu_0} \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{\phi}_0) \\ &\propto g(\boldsymbol{\theta})^{\nu_0 + n} \exp\left[\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} \left(\boldsymbol{\phi}_0 + \textstyle\sum_{i=1}^{n} t(\boldsymbol{y}_i) \right) \right] \\ &\propto \pi\left(\boldsymbol{\theta}; \nu_0 + n, \boldsymbol{\phi}_0 + \textstyle\sum_{i=1}^{n} t(\boldsymbol{y}_i) \right). \end{aligned}$$

The form of the update equations leads to an interpretation of the prior hyperparameters ν_0 and $\boldsymbol{\phi}_0$. The prior distribution $pi(\boldsymbol{\theta}; \nu_0, \boldsymbol{\phi}_0)$ acts as a pseudo-dataset of ν_0 observations with sufficient statistics $\boldsymbol{\phi}_0$.

In practice it is common to use more flexible priors than permitted by the standard conjugate family (Gutiérrez-Peña et al., 1997; Arnold et al., 1993). For example, suppose we have n observations from a d-dimensional multivariate normal distribution with known covariance matrix Σ and unknown mean $\boldsymbol{\mu}$. Using definition (2.49), the standard conjugate family has prior $N(\boldsymbol{m}_0, \Sigma/\nu_0)$ for some prior mean $\boldsymbol{m}_0 \in \mathbb{R}^d$ and positive scalar ν_0. The prior covariance of $\boldsymbol{\mu}$ must be proportional to the covariance matrix of the observations. In practice, a common conjugate prior on $\boldsymbol{\mu}$ is a $N(\boldsymbol{m}_0, \boldsymbol{V}_0)$ distribution (Gelman et al., 2014) where $\boldsymbol{V}_0$ is any positive definite $d \times d$ matrix. The flexibility of an unstructured prior covariance matrix $\boldsymbol{V}_0$ places it outside the standard class (2.49).

An enriched conjugate family of priors introduces an extra r-dimensional hyperparameter $\boldsymbol{\omega}_0 = (\omega_1, \ldots, \omega_r)$, and r positive valued functions $b_u(\boldsymbol{\theta}) : \Omega \to \mathbb{R}^+$ for $u = 1, \ldots, r$ (Gutiérrez-Peña et al., 1997). Define the function

$$b(\boldsymbol{\theta}|\boldsymbol{\omega}_0) = \prod_{h=1}^{r} [b_u(\boldsymbol{\theta})]^{\omega_u}. \tag{2.50}$$

The enriched conjugate prior is defined as

$$\pi(\boldsymbol{\theta}; \nu_0, \boldsymbol{\phi}_0, \omega_0) = c(\nu_0, \boldsymbol{\phi}_0, \boldsymbol{\omega}_0) b(\boldsymbol{\theta}|\boldsymbol{\omega}_0) g(\boldsymbol{\theta})^{\nu_0} \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{\phi}_0). \tag{2.51}$$

In the previous display $c(\nu_0, \boldsymbol{\phi}_0, \boldsymbol{\omega}_0)$ gives the normalising constant

$$\frac{1}{c(\nu_0, \boldsymbol{\phi}_0, \boldsymbol{\omega}_0)} = \int b(\boldsymbol{\theta}|\boldsymbol{\omega}_0) g(\boldsymbol{\theta})^{\nu_0} \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{\phi}_0) \, d\boldsymbol{\theta}.$$

Standard conjugate priors (2.49) can be viewed as a special case of the enriched form (2.51) where $\boldsymbol{\omega}_0 = \mathbf{0}$, and the functions $b_1(\boldsymbol{\theta}), \ldots, b_r(\boldsymbol{\theta})$ can be set to return one for all $\boldsymbol{\theta}$. The update equations for the enriched prior are very similar to the update equations for the standard prior. It is simple to show that if we have n independent and identically distributed observations from model (2.48), the posterior remains in the enriched family,

$$p(\boldsymbol{\theta}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \pi\left(\boldsymbol{\theta}; \nu_0 + n, \boldsymbol{\phi}_0 + \textstyle\sum_{i=1}^{n} t(\boldsymbol{y}_i), \boldsymbol{\omega}_0\right).$$

The ω_0 hyperparameter remains unchanged, and the data still influences the posterior through the term $\boldsymbol{\phi}_0 + \sum_{i=1}^{n} t(\boldsymbol{y}_i)$. Almost all commonly used conjugate priors can be shown to be in the enriched family (Arnold et al., 1993). However, it can be tedious to reparametrise benchmark priors to match the enriched form. It can be shown that the $N(\boldsymbol{\mu}_0, \boldsymbol{V}_0)$ prior belongs to the enriched conjugate prior family for Gaussian likelihoods (Gutiérrez-Peña et al., 1997, Example 4.3). The abstraction of the enriched conjugate prior family is beneficial as it provides a unifying framework for establishing general results.

2.5.3 Data augmentation

It is not possible to write the likelihood function in the form (2.48) for many interesting statistical models. Data augmentation is a useful technique that extends the utility of exponential family and conjugate theory to a wider class of models. Suppose that we have an observed dataset of n observations $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$. A wide range of data augmentation strategies add a hidden layer of latent variables $\boldsymbol{z}$ with state space $\mathcal{Z}$ to the model, such that the observed data likelihood satisfies

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \int_{\mathcal{Z}} p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta}) \, d\boldsymbol{z}.$$

The latent variables $\boldsymbol{z}$ often have a natural statistical interpretation in terms of a hidden layer in the data generating process. Ideally, the complete data likelihood $p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta})$ is then more tractable than the observed data likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$. In the simplest form, we introduce n independent latent variable $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$, one for each observation in the original dataset. With n independent latent variables

and n independent observations, the complete data likelihood satisfies

$$p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{y}_i, \boldsymbol{z}_i|\boldsymbol{\theta}).$$

Suppose that the complete data likelihood contribution for observation i belongs to the exponential family, so

$$p(\boldsymbol{y}_i, \boldsymbol{z}_i|\boldsymbol{\theta}) = h(\boldsymbol{y}_i, \boldsymbol{z}_i)g(\boldsymbol{\theta}) \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} t(\boldsymbol{y}_i, \boldsymbol{z}_i)), \tag{2.52}$$

again for some known functions $h(\cdot), g(\cdot), t(\cdot)$ and $\boldsymbol{\eta}(\cdot)$. Suppose that we adopt an enriched conditionally conjugate prior $p(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}; \nu_0, \boldsymbol{\phi}_0, \boldsymbol{\omega}_0)$. It is immediate that the conditional posterior has the form

$$p(\boldsymbol{\theta}|\boldsymbol{z}, \boldsymbol{y}) = \pi(\boldsymbol{\theta}; \nu_0 + n, \boldsymbol{\phi}_0 + \textstyle\sum_{i=1}^{n} t(\boldsymbol{y}_i, \boldsymbol{z}_i), \boldsymbol{\omega}_0). \tag{2.53}$$

The conditional posterior depends on the sufficient statistics of the augmented dataset $\sum_{i=1}^{n} t(\boldsymbol{y}_i, \boldsymbol{z}_i)$. For many models, it is possible to sample from the full conditional of the latent variables $p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{y})$. The joint posterior on the unknowns $p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{y})$ can thus be targeted using a two block Gibbs sampler, iteratively sampling from the full conditionals $p(\boldsymbol{\theta}|\boldsymbol{z}, \boldsymbol{\theta})$ and $p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{y})$. This data augmentation and Gibbs approach can be used for logistic regression, negative binomial regression, probit regression and Gaussian mixtures.

2.5.4 Chib's method

Chib's method for marginal likelihood computation (Chib, 1995) using Gibbs sampling is based on the integrated likelihood identity (2.16). The method uses the structure of data augmented models to estimate the pointwise density in a more efficient manner than nonparametric kernel density estimation. As mentioned in section 2.5.2, it is often possible to introduce a vector of latent variables $\boldsymbol{z}$ such that the conditional posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z})$ is known exactly. Marginalising over the posterior distribution of the latent variables $\boldsymbol{z}$ gives the relationship

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \int p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z}) p(\boldsymbol{z}|\boldsymbol{y}). \tag{2.54}$$

A simulation consistent estimate of the integrated likelihood can be obtained by simulating from the posterior distribution of the latent variables $\boldsymbol{z}$. As discussed, it is common to use a two block Gibbs sampler, iteratively sampling the full conditional distributions $p(\boldsymbol{\theta}|\boldsymbol{z}, \boldsymbol{y})$ and $p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{y})$. Suppose we use a conditionally conjugate prior, so the full conditional $p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z})$ is known explicitly. Given B posterior samples of the latent variables, $\boldsymbol{z}^{[1]}, \ldots, \boldsymbol{z}^{[b]}$, a simulation consistent estimator of the posterior density at $\boldsymbol{\theta}^* \in \Omega$ is then

$$\widehat{p}(\boldsymbol{\theta}^*|\boldsymbol{y}) = \frac{1}{B} \sum_{b=1}^{B} p(\boldsymbol{\theta}^*|\boldsymbol{z}^{[b]}, \boldsymbol{y}). \tag{2.55}$$

Substituting (2.55) into the integrated likelihood identity (2.16) gives a simulation consistent estimator of the integrated likelihood,

$$\log \widehat{p}(\boldsymbol{y}) = \log p(\boldsymbol{y}|\boldsymbol{\theta}^*) + \log p(\boldsymbol{\theta}^*) - \log \left(\frac{1}{B} \sum_{b=1}^{B} p(\boldsymbol{\theta}^*|\boldsymbol{y}, \boldsymbol{z}^{[b]}) \right). \tag{2.56}$$

Chib's method can be used for high-dimensional models where general non-parametric density estimators would likely fail. The key to the approach is the marginal representation of the posterior density (2.55). We can take an average over posterior draws of the latent variables to estimate the unknown posterior density $p(\boldsymbol{\theta}^*|\boldsymbol{y})$. In the following subsections we develop an embarrassingly parallel algorithm that makes use of the fact that although

$$\int \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i) \, d\boldsymbol{\theta},$$

may be intractable, data augmentation with suitably chosen latent variables in each subset $\boldsymbol{z}_i$ could lead to the augmented subposterior integral

$$\int \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{z}_i)\ d\boldsymbol{\theta},$$

having a closed form solution.

2.5.5 Apply step

We first show how data augmentation and Gibbs sampling can be used to draw subposterior samples and compute the subposterior evidence scores in the apply step. We assume that the complete data likelihood belongs to the exponential family. Let n_i give the number of observations in subset i for $i = 1, \ldots, s$. Let $\boldsymbol{y}_{ij}$ denote the jth observation in subset i, and let $\boldsymbol{z}_{ij}$ denote an associated latent variable for the jth observation in subset i. We assume that the latent variables $\boldsymbol{z}_{ij}$ are independent given $\boldsymbol{\theta}$. Finally let $\boldsymbol{z}_i$ give the vector of latent variables for subset i, that is $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{in_i})$ for $i = 1, \ldots, s$. The augmented full dataset posterior distribution is again proportional to s subposteriors:

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_s) &\propto p(\boldsymbol{\theta}) p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_s|\boldsymbol{\theta}) \\ &= \prod_{i=1}^{s} p(\boldsymbol{y}_i, \boldsymbol{z}_i|\boldsymbol{\theta}) p(\boldsymbol{\theta})^{1/s} \\ &\propto \prod_{i=1}^{s} p(\boldsymbol{y}_i, \boldsymbol{z}_i|\boldsymbol{\theta}) \widetilde{p}(\boldsymbol{\theta}) \\ &\propto \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{z}_i). \end{aligned}$$

Suppose the complete data likelihood can be written in the following form

$$p(\boldsymbol{y}_{ij}, \boldsymbol{z}_{ij}|\boldsymbol{\theta}) = h(\boldsymbol{y}_{ij}, \boldsymbol{z}_{ij}) g(\boldsymbol{\theta}) \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} t(\boldsymbol{y}_{ij}, \boldsymbol{z}_{ij})). \tag{2.57}$$

Where the functions $h(\cdot), g(\cdot), t(\cdot)$ and $\boldsymbol{\eta}(\cdot)$ are known. Suppose the original prior is of the enriched conjugate form (2.51). The fractionated prior remains in the enriched conjugate family. Fractionating yields

$$\begin{aligned} p(\boldsymbol{\theta}|\nu_0, \boldsymbol{\phi}_0, \boldsymbol{\omega}_0)^{1/s} &= \left[c(\nu_0, \boldsymbol{\phi}_0, \boldsymbol{\omega}_0) b(\boldsymbol{\theta}|\boldsymbol{\omega}_0) g(\boldsymbol{\theta})^{\nu_0} \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{\phi}_0)\right]^{1/s} \\ &= c(\nu_0, \boldsymbol{\phi}_0, \boldsymbol{\omega}_0)^{1/s} b(\boldsymbol{\theta}|\boldsymbol{\omega}_0)^{1/s} g(\boldsymbol{\theta})^{\nu_0/s} \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{\phi}_0/s). \end{aligned}$$

Recall that $\boldsymbol{\omega}_0 = (\omega_1, \ldots, \omega_r)$ and the definition of the function $b(\boldsymbol{\theta}|\boldsymbol{\omega}_0) = \prod_{u=1}^{r} [b_u(\boldsymbol{\theta})]^{\omega_u}$. It then follows that $b(\boldsymbol{\theta}|\boldsymbol{\omega}_0)^{1/s} = b(\boldsymbol{\theta}|\boldsymbol{\omega}_0/s)$. Continuing, we see that the fractionated prior remains in the enriched family

$$\begin{aligned} p(\boldsymbol{\theta}|\nu_0, \boldsymbol{\phi}_0, \boldsymbol{\omega}_0)^{1/s} &= c(\nu_0, \boldsymbol{\phi}_0, \boldsymbol{\omega}_0)^{1/s} b(\boldsymbol{\theta}|\boldsymbol{\omega}_0/s) g(\boldsymbol{\theta})^{\nu_0/s} \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{\phi}_0/s) \\ &\propto \pi(\boldsymbol{\theta}; \nu_0/s, \boldsymbol{\phi}_0/s, \boldsymbol{\omega}_0/s). \end{aligned}$$

Given the original prior is $p(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}; \nu_0, \boldsymbol{\phi}_0, \boldsymbol{\omega}_0)$, the subprior can be defined as $\widetilde{p}(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}; \nu_0/s, \boldsymbol{\phi}_0/s, \boldsymbol{\omega}_0/s)$. The hyperparameters of the subprior show that each subset analysis receives a fraction of the original prior pseudo-dataset, namely ν_0/s observations with sufficient statistics $\boldsymbol{\phi}_0/s$. Secondly, as the $\boldsymbol{\omega}_0$ hyperparameter tends to $\mathbf{0}$ the function $b(\boldsymbol{\theta}|\boldsymbol{\omega})$ tends to one. The subpriors have hyperparameters $\boldsymbol{\omega}_0/s$ which indicates a dampened effect of the extra prior flexibility function $b(\boldsymbol{\theta}|\boldsymbol{\omega})$. There are open questions surrounding the importance of prior fractionation (Scott, 2017; Gelman and Vehtari, 2017), and the theory of exponential families can be of use in addressing these.

From the results in section 2.5.3, it follows that given the latent variables $\boldsymbol{z}_i$, the subposteriors $\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{z}_i)$ are also conditionally conjugate. For $i = 1, \ldots, s$:

$$\begin{aligned}\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{z}_i) &\propto g(\boldsymbol{\theta})^{n_i} \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} \textstyle\sum_{i=j}^{n_i} t(\boldsymbol{y}_{ij}, \boldsymbol{z}_{ij})\right) p(\boldsymbol{\theta}|\nu_0, \boldsymbol{\phi}_0, \boldsymbol{\omega}_0)^{1/s} \\ &\propto g(\boldsymbol{\theta})^{n_i} \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} \textstyle\sum_{j=1}^{n_i} t(\boldsymbol{y}_{ij}, \boldsymbol{z}_{ij})\right) \pi(\boldsymbol{\theta}; \nu_0/s, \boldsymbol{\phi}_0/s, \boldsymbol{\omega}_0/s) \\ &\propto \pi\left(\boldsymbol{\theta} \,\middle|\, \nu_0/s + n_i, \boldsymbol{\phi}_0/s + \textstyle\sum_{j=1}^{n_i} t(\boldsymbol{y}_{ij}, \boldsymbol{z}_{ij}), \boldsymbol{\omega}_0/s\right).\end{aligned}$$

Suppose that it is possible to sample from the full conditional $p(\boldsymbol{z}_i|\boldsymbol{\theta}, \boldsymbol{y}_i)$ in a single machine analysis. It is then also possible to sample from conditional subposterior distribution $\widetilde{p}(\boldsymbol{z}_i|\boldsymbol{\theta}, \boldsymbol{y}_i)$ using the same update equations. The conditional subposterior has the same form as in regular analysis of a small portion of the full dataset:

$$\begin{aligned}\widetilde{p}(\boldsymbol{z}_i|\boldsymbol{y}_i, \boldsymbol{\theta}) &\propto p(\boldsymbol{y}_i, \boldsymbol{z}_i|\boldsymbol{\theta})\widetilde{p}(\boldsymbol{\theta}) \\ &\propto p(\boldsymbol{z}_i|\boldsymbol{\theta}, \boldsymbol{y}_i)p(\boldsymbol{y}_i|\boldsymbol{\theta}) \\ &\propto p(\boldsymbol{z}_i|\boldsymbol{\theta}, \boldsymbol{y}_i).\end{aligned}$$

Use of the subprior $\widetilde{p}(\boldsymbol{\theta})$ is not material when conditioning on $\boldsymbol{\theta}$. We keep the tilde to acknowledge that both $p(\boldsymbol{z}_i|\boldsymbol{\theta}, \boldsymbol{y}_i)$ and $\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{z}_i)$ are targeted in the apply stage by an invidual worker. As such, we can use Gibbs sampling to target the subposterior distribution $\widetilde{p}(\boldsymbol{\theta}, \boldsymbol{z}_i|\boldsymbol{y}_i)$ by sampling from the conditional densities $\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{z}_i)$ and $\widetilde{p}(\boldsymbol{z}_i|\boldsymbol{\theta}, \boldsymbol{y}_i)$. We have previously assumed that each worker returns the subposterior evidence scores. We can use Chib's method in the apply stage to compute the subposterior evidence values. Let $\boldsymbol{z}_i^{[b]}$ be the sampled latent variables at iteration b in subposterior i. Assuming the chain has mixed well, $\boldsymbol{z}_i^{[b]}$ is a sample from the distribution $\widetilde{p}(\boldsymbol{z}_i|\boldsymbol{y}_i)$. At any particular ordinate $\boldsymbol{\theta}^* \in \Omega$ we have the identity

$$\log\ \widetilde{p}(\boldsymbol{y}_i) = \log p(\boldsymbol{y}_i|\boldsymbol{\theta}^*) + \log \widetilde{p}(\boldsymbol{\theta}^*) - \log \widetilde{p}(\boldsymbol{\theta}^*|\boldsymbol{y}_i),$$

A simulation consistent estimator of $\log\ \widetilde{p}(\boldsymbol{y}_i)$ is therefore

$$\widehat{\log}\ \widetilde{p}(\boldsymbol{y}_i) = \log p(\boldsymbol{y}_i|\boldsymbol{\theta}^*) + \log \widetilde{p}(\boldsymbol{\theta}^*) - \log\left(\frac{1}{B}\sum_{b=1}^{B} \widetilde{p}(\boldsymbol{\theta}^*|\boldsymbol{y}_i, \boldsymbol{z}_i^{[b]})\right). \tag{2.58}$$

2.5.6 Combine step

The sequence of full conditional Gibbs distributions in the apply stage can be used to give a closed form estimator of the subposterior integral

$$I_{\text{sub}} = \int \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i)\ d\boldsymbol{\theta}. \tag{2.59}$$

To describe the estimator of I_{sub} suppose the latent variables $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s$ take values in state spaces $\mathcal{Z}_1, \ldots, \mathcal{Z}_s$ respectively. Let the joint space of the subposterior latent variables be denoted $\mathcal{Z} = \mathcal{Z}_1 \times \ldots \times \mathcal{Z}_s$. Each subposterior has the marginal representation $\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i) = \int_{\mathcal{Z}_i} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{z}_i)\widetilde{p}(\boldsymbol{z}_i|\boldsymbol{y}_i)\ d\boldsymbol{z}_i$. The subposterior latent variables can be considered together as a random vector $(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_s)$ with joint distribution $\widetilde{p}(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_s) = \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{z}_i|\boldsymbol{y}_i)$. As such, we have the following representation of the product of the subposterior distributions

$$\begin{aligned}\prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i) &= \prod_{i=1}^{s} \left(\int_{\mathcal{Z}_i} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{z}_i)\widetilde{p}(\boldsymbol{z}_i|\boldsymbol{y}_i)\ d\boldsymbol{z}_i\right) \\ &= \int_{\mathcal{Z}_1 \times \ldots \times \mathcal{Z}_s} \left(\prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{z}_i)\widetilde{p}(\boldsymbol{z}_i|\boldsymbol{y}_i)\right) d\boldsymbol{z}_1 d\boldsymbol{z}_2 \ldots d\boldsymbol{z}_s \\ &= \int_{\mathcal{Z}} \left(\prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{z}_i)\right)\left(\prod_{i=1}^{s} \widetilde{p}(\boldsymbol{z}_i|\boldsymbol{y}_i)\right) d\boldsymbol{z}_1 d\boldsymbol{z}_2 \ldots d\boldsymbol{z}_s.\end{aligned} \tag{2.60}$$

Switching the integration and product operators is justified using Fubini's theorem. The necessary assumptions are satisfied if all the subposterior distributions are proper (Keener, 2013). Substituting (2.60) into the definition of I_{sub} (2.59) gives

$$\int_\Omega \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i)\ d\boldsymbol{\theta} = \int_\Omega \int_{\mathcal{Z}} \left(\prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{z}_i)\right) \left(\prod_{i=1}^{s} \widetilde{p}(\boldsymbol{z}_i|\boldsymbol{y}_i)\right) d\boldsymbol{z}_1 d\boldsymbol{z}_2 \ldots d\boldsymbol{z}_s\ d\boldsymbol{\theta}$$

Using Fubini's theorem once more allows the order of integration to be switched,

$$\int_\Omega \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i)\ d\boldsymbol{\theta} = \int_{\mathcal{Z}} \left(\int_\Omega \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{z}_i)\ d\boldsymbol{\theta}\right) \left(\prod_{i=1}^{s} \widetilde{p}(\boldsymbol{z}_i|\boldsymbol{y}_i)\right)\ d\boldsymbol{z}_1 d\boldsymbol{z}_2 \ldots d\boldsymbol{z}_s \tag{2.61}$$

The inner integral over Ω can be obtained in closed form given that the subposteriors are conditionally conjugate. The subposterior integral is equal to the expected value of the augmented subposterior integral, where the latent variables $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s$ are sampled from the subposterior distributions. We can write

$$I_{\text{sub}} = \mathbb{E}_{\widetilde{p}(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_s|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s)} \left[\int_\Omega \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{z}_i)\ d\boldsymbol{\theta}\right] \tag{2.62}$$

For convenience, let $t(\boldsymbol{y}_i, \boldsymbol{z}_i)$ denote the complete data sufficient statistic for subset i, that is $t(\boldsymbol{y}_i, \boldsymbol{z}_i) = \sum_{j=1}^{n_i} t(\boldsymbol{y}_{ij}, \boldsymbol{z}_{ij})$. For more compact notation define for $i = 1, \ldots, s$ the subset normalising constants

$$c_i = c(\nu_0/s + n_i, \boldsymbol{\phi}_0/s + t(\boldsymbol{y}_i, \boldsymbol{z}_i), \boldsymbol{\omega}_0/s),$$

where $c(\cdot)$ is the normalising function for the enriched conjugate prior. Additionally, define the full dataset normalising constant as

$$C = c(\nu_0 + n, \boldsymbol{\phi}_0 + \textstyle\sum_{i=1}^{s} t(\boldsymbol{y}_i, \boldsymbol{z}_i), \boldsymbol{\omega}_0).$$

Using properties of the standard conjugate prior we obtain a closed form expression for the augmented subposterior integral.

$$\begin{aligned}
\int_\Omega \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{z}_i)\ d\boldsymbol{\theta} &= \int_\Omega \prod_{i=1}^{s} c_i \times b(\boldsymbol{\theta}|\boldsymbol{\omega}_0/s) g(\boldsymbol{\theta})^{\nu_0/s+n_i} \exp\left[\boldsymbol{\eta}(\boldsymbol{\theta})^\mathsf{T} \left(\boldsymbol{\phi}_0/s + t(\boldsymbol{y}_i, \boldsymbol{z}_i)\right)\right]\ d\boldsymbol{\theta} \\
&= \left(\prod_{i=1}^{s} c_i\right) \int_\Omega g(\boldsymbol{\theta})^{\nu_0+n} b(\boldsymbol{\theta}|\boldsymbol{\omega}_0) \exp\left[\boldsymbol{\eta}(\boldsymbol{\theta})^\mathsf{T} \left(\boldsymbol{\phi}_0 + \sum_{i=1}^{s} t(\boldsymbol{y}_i, \boldsymbol{z}_i)\right)\right]\ d\boldsymbol{\theta} \\
&= \left(\frac{\prod_{i=1}^{s} c_i}{C}\right) \int_\Omega p(\boldsymbol{\theta}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_s)\ d\boldsymbol{\theta} \\
&= \frac{\prod_{i=1}^{s} c(\nu_0/s + n_i, \boldsymbol{\phi}_0/s + t(\boldsymbol{y}_i, \boldsymbol{z}_i), \boldsymbol{\omega}_0/s)}{c(\nu_0 + n, \boldsymbol{\phi}_0 + \sum_{i=1}^{s} t(\boldsymbol{y}_i, \boldsymbol{z}_i), \boldsymbol{\omega}_0)}.
\end{aligned} \tag{2.63}$$

The second line uses the fact that $\prod_{i=1}^{s} b(\boldsymbol{\theta}|\boldsymbol{\omega}_0/s) = b(\boldsymbol{\theta}|\boldsymbol{\omega}_0)$. Substituting (2.63) into (2.64) gives an important expression for the subposterior integral I_{sub}.

$$I_{\text{sub}} = \mathbb{E}_{\widetilde{p}(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_s|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s)} \left[\frac{\prod_{i=1}^{s} c(\nu_0/s + n_i, \boldsymbol{\phi}_0/s + t(\boldsymbol{y}_i, \boldsymbol{z}_i), \boldsymbol{\omega}_0/s)}{c(\nu_0 + n, \boldsymbol{\phi}_0 + \sum_{i=1}^{s} t(\boldsymbol{y}_i, \boldsymbol{z}_i), \boldsymbol{\omega}_0)}\right] \tag{2.64}$$

Let $\boldsymbol{\phi}_i^{[b]} = \boldsymbol{\phi}_0/s + t(\boldsymbol{z}_i^{[b]}, \boldsymbol{y}_i)$ be the data dependent parameter of the conditional subposterior at iteration b for $i = 1, \ldots, s$. Suppose we have saved either the sufficient statistics $t(\boldsymbol{y}_i, \boldsymbol{z}_i^{[b]})$ or conditional posterior parameters $\boldsymbol{\phi}_i^{[b]}$ at each iteration $b = 1, \ldots, B$ in each minibatch analysis $i = 1, \ldots, s$. The subposterior integral can then be estimated in the combine step by pooling the Gibbs histories from the apply step.

In full,

$$\begin{aligned}\widehat{I}_{\text{sub}} &= \frac{1}{B}\sum_{b=1}^{B}\int_{\Omega}\prod_{i=1}^{s}\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i,\boldsymbol{z}_i^{[b]})\,d\boldsymbol{\theta} \\ &= \frac{1}{B}\sum_{b=1}^{B}\frac{\prod_{i=1}^{s}c(\nu_0/s+n_i,\boldsymbol{\phi}_0/s+t(\boldsymbol{y}_i,\boldsymbol{z}_i^{[b]}))}{c\left(\nu_0+n,\boldsymbol{\phi}_0+\sum_{i=1}^{s}t(\boldsymbol{z}_i^{[b]},\boldsymbol{y}_i)\right)} \qquad (2.65)\\ &= \frac{1}{B}\sum_{b=1}^{B}\frac{\prod_{i=1}^{s}c(\nu_0/s+n_i,\boldsymbol{\phi}_i^{[b]})}{c(\nu_0+n,\sum_{i=1}^{s}\boldsymbol{\phi}_i^{[b]})}. \qquad (2.66)\end{aligned}$$

Equations (2.65) and (2.66) both define plug-in Monte Carlo estimators of the expectation in equation (2.64).

2.5.7 Embarrassingly parallel evidence estimation

Recall (2.27), the subset decomposition of the full dataset model evidence

$$\log p(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s) = \left(\sum_{i=1}^{s}\log\widetilde{p}(\boldsymbol{y}_i)\right) + \log I_{\text{sub}} + s\log\alpha.$$

The sum over the subposterior evidence scores can be estimated during the apply stage using Chib's method. To estimate the sequence of full conditional distributions in the combine stage it is necessary to save information about the full conditional distributions that are sampled from in the apply stage. The subprior normalising constant α can generally be determined in closed form given that the original prior is in the exponential familty. The full dataset model evidence is then estimated by combining the subset evidence scores with the subposterior integral,

$$\widehat{\log}\ p(\boldsymbol{y}_i,\ldots,\boldsymbol{y}_s) = \left(\sum_{i=1}^{s}\widehat{\log}\ \widetilde{p}(\boldsymbol{y}_i)\right) + \log\widehat{I}_{\text{sub}} + s\log\alpha.$$

Algorithm 2.2 presents the proposed methodology in detail.

Algorithm 2.2 Divide and conquer for conditionally conjugate models

Apply Step: Subset Markov chain Monte Carlo (MCMC) runs
for $i = 1,\ldots,s$ **do**
  for $b = 1,\ldots,B$ **do**
    Sample $\boldsymbol{z}_i^{[b]} \sim p(\boldsymbol{z}_i|\boldsymbol{\theta},\boldsymbol{y}_i)$
    Compute $\boldsymbol{\phi}_i^{[b]} = \boldsymbol{\phi}_0/s + t(\boldsymbol{z}_i^{[b]},\boldsymbol{y}_i)$
    Save $\boldsymbol{\phi}_i^{[b]}$
    Sample $\boldsymbol{\theta}$ from $\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{z}_i^{[b]},\boldsymbol{y}_i) = \pi(\boldsymbol{\theta};\nu_0/s+n_i,\boldsymbol{\phi}_i^{[b]},\boldsymbol{\omega}_0/s)$
  end for
  Compute $\widehat{\log}\ \widetilde{p}(\boldsymbol{y}_i)$ using (2.58) or alternative method.
end for

Combine Step: Post processing subset Gibbs output
for $b = 1,\ldots,B$ **do**
  Compute $g_b = \dfrac{\prod_{i=1}^{s}c(\nu_0/s,\boldsymbol{\phi}_i^{[b]},\boldsymbol{\omega}_0/s)}{c(\nu_0,\sum_{i=1}^{s}\boldsymbol{\phi}_i^{[b]},\boldsymbol{\omega}_0)}$
end for
Compute $\widehat{I}_{\text{sub}} = B^{-1}\sum_{b=1}^{B}g_b$
Compute $\widehat{\log}\ p(\boldsymbol{y}_i,\ldots,\boldsymbol{y}_s) = \left(\sum_{i=1}^{s}\widehat{\log}\ \widetilde{p}(\boldsymbol{y}_i)\right) + \log\widehat{I}_{\text{sub}} + s\log\alpha$

To recap, we have developed an embarrassingly parallel algorithm to compute the full dataset model evidence using distributed computing and Gibbs sampling. The algorithm meets the desired template of Figure 2.1. The complexity of the combine step does not depend on n. It is quite remarkable that we can

estimate the integrated likelihood on massive datasets without ever fitting the model to the whole dataset. This is a significant departure from many other techniques for computing the integrated likelihood, where generating samples from the full dataset posterior $p(\boldsymbol{\theta}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)$ is an important intermediate step (Friel and Wyse, 2012). Existing methods for calculating the integrated likelihood are discussed in more detail in Chapter 3.

2.5.8 Monte Carlo error

For Algorithm 2.2 to be effective, the Monte Carlo error in the combine step needs to be tolerable. Algorithm 2.1 is impractical for this reason, due to the curse of dimensionality impeding nonparametric kernel density estimation of the subposterior integral. In this section we show that the variance of the Gibbs based estimatpor subposterior integral (2.66) is tied to the disparity between the subposterior distribution on the latent variables, and the full dataset posterior distribution. The variance can be characterised in terms of importance sampling. As a quick review, importance sampling is a basic Monte Carlo technique based on a change of measure. Suppose we are interested in the expectation of a function $h(\boldsymbol{x})$ over some probability distribution $p(\boldsymbol{x})$ on state space $\mathcal{X}$. Importance sampling is based on the simple identity,

$$\int_{\mathcal{X}} h(\boldsymbol{x}) p(\boldsymbol{x}) \, d\boldsymbol{x} = \int_{\mathcal{X}} h(\boldsymbol{x}) \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} q(\boldsymbol{x}) \, d\boldsymbol{x},$$

for some other distribution $q(\boldsymbol{x})$ on the same space $\mathcal{X}$. A common use for importance sampling is when $p(\boldsymbol{x})$ is difficult to sample from, but it is easy to find an approximating distribution $q(\boldsymbol{x})$. An important diagnostic in importance sampling is the variance of the importance sampling weights

$$\mathrm{var}_{q(\boldsymbol{x})} \left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \right).$$

If $q(\boldsymbol{x}) \propto p(\boldsymbol{x})$, then the variance of the weights is zero. The variance of the weights is strongly influenced by the differences between $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$. The tails of the distribution $q(\boldsymbol{x})$ are particularly significant. If $p(\boldsymbol{x})$ has high support in regions that are minimally covered by $q(\boldsymbol{x})$, the variance of the importance weights will be high, and the importance sampler will be ineffective. It is possible for the variance of the weights to be infinite, in which case there can be a large finite sample bias on the importance sampling estimator (Robert and Casella, 2010).

Let $\widetilde{p}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)$ be the effective joint distribution on the latent variables if sampling independently from each subposterior, so $\widetilde{p}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s) = \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{z}_i|\boldsymbol{y}_i)$. The target posterior on the latent variables is $p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)$. To derive the variance of $\widehat{I}_{\mathrm{sub}}$ we start by applying Bayes theorem to the augmented posterior

$$\begin{aligned}
p(\boldsymbol{\theta}|\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_s) &= \frac{p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_s|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)} \\
&= \frac{\alpha^s}{p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s) p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)} \prod_{i=1}^{s} p(\boldsymbol{y}_i, \boldsymbol{z}_i|\boldsymbol{\theta}) \widetilde{p}(\boldsymbol{\theta}) \\
&= \frac{\alpha^s}{p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s) p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)} \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{z}_i) \widetilde{p}(\boldsymbol{z}_i|\boldsymbol{y}_i) \widetilde{p}(\boldsymbol{y}_i) \\
&= \frac{\alpha^s \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{y}_i)}{p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)} \frac{\prod_{i=1}^{s} \widetilde{p}(\boldsymbol{z}_i|\boldsymbol{y}_i)}{p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)} \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{z}_i). \qquad (2.67)
\end{aligned}$$

The core identity

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s) = \left(\prod_{i=1}^{s} \widetilde{p}(\boldsymbol{y}_i) \right) \alpha^s \times I_{\mathrm{sub}}$$

gives the relationship

$$\frac{\prod_{i=1}^{s}\widetilde{p}(\boldsymbol{y}_i)}{p(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s)} = \frac{\alpha^{-s}}{I_{\text{sub}}}. \tag{2.68}$$

Substituting (2.68) into (2.67) and recalling the definition $\widetilde{p}(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_s|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s) = \prod_{i=1}^{s}\widetilde{p}(\boldsymbol{z}_i|\boldsymbol{y}_i)$,

$$p(\boldsymbol{\theta}|\boldsymbol{z}_1,\ldots,\boldsymbol{z}_s,\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s) = \frac{\alpha^s\alpha^{-s}}{I_{\text{sub}}}\frac{\widetilde{p}(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_s|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s)}{p(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_s|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s)}\prod_{i=1}^{s}\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i,\boldsymbol{z}_i).$$

Integrating both sides over $\boldsymbol{\theta}$ gives

$$1 = \frac{1}{I_{\text{sub}}}\frac{\widetilde{p}(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_s|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s)}{p(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_s|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s)}\int\prod_{i=1}^{s}\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i,\boldsymbol{z}_i)\;d\boldsymbol{\theta}.$$

Rearranging shows a connection between the integral over the augmented subposteriors and importance sampling weights.

$$\int\prod_{i=1}^{s}\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i,\boldsymbol{z}_i)\;d\boldsymbol{\theta} = I_{\text{sub}}\times\frac{p(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_s|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s)}{\widetilde{p}(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_s|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s)} \tag{2.69}$$

The subposterior integral estimator is defined as the plug-in expectation of the augmented integral over $\widetilde{p}(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_s|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s)$. The estimator can be written as

$$\begin{aligned}\widehat{I}_{\text{sub}} &= \frac{1}{B}\sum_{b=1}^{B}\int_{\Omega}\prod_{i=1}^{s}\widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i,\boldsymbol{z}_i^{[b]})\;d\boldsymbol{\theta}\\ &= \frac{1}{B}\times I_{\text{sub}}\times\sum_{b=1}^{B}\frac{p(\boldsymbol{z}_1^{[b]},\ldots,\boldsymbol{z}_s^{[b]}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s)}{\widetilde{p}(\boldsymbol{z}_1^{[b]},\ldots,\boldsymbol{z}_s^{[b]}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s)}.\end{aligned}$$

Assuming independent subposterior samples, the variance of the estimator is therefore

$$\text{var}(\widehat{I}_{\text{sub}}) = \frac{1}{B}\times I_{\text{sub}}^2\times\text{var}_{\widetilde{p}(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_s|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s)}\left(\frac{p(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_s|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s)}{\widetilde{p}(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_s|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s)}\right). \tag{2.70}$$

In reality, correlated subposterior samples will inflate the variance, but to simplify the discussion we assume independence or samples from a sufficiently thinned Markov chain. The variance of $\widehat{I}_{\text{sub}}$ is tied to the mean multiplied by a density ratio involving the latent variables in the model. The final term in brackets can be interpreted as importance sampling weights if using the subposterior distributions over the latent variables to approximate the full dataset joint posterior on the latent variables. There are two important conclusions. The first is that the variance of the integral estimator is only finite if the variance of the importance sampling weights is finite. If the tails of the subposterior distribution $\widetilde{p}(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_s|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s)$ contain regions where the target posterior $p(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_s|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s)$ has high support, then the estimator $\widehat{I}_{\text{sub}}$ could have infinite variance. The second conclusion in the converse direction is that it is possible to obtain a zero variance estimator if $\widetilde{p}(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_s|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s)$ is proportional to $p(\boldsymbol{z}_1,\ldots,\boldsymbol{z}_s|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_s)$.

It is difficult to make a broad statement about how the variance of the proposed estimator (2.66) scales with the number of subsets s. The variance is tied to the model, the goodness of fit of the model and how the full dataset is partitioned. However, Algorithm 2.2 has more promise than the approach based on kernel density estimation as the curse of dimensionality does not completely rule out its efficacy.

2.6 Logistic regression

We now show how Algorithm 2.2 can be used for large scale logistic regression. Suppose we have n binary observations $y_i \in \{0,1\}$ with an associated p-dimensional vector of covariates $\boldsymbol{x}_i$ for $i = 1,\ldots,n$. The responses are modelled as $y_i \sim \text{Bernoulli}(\sigma(\eta_i))$, where $\eta_i = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}$ and $\sigma(\eta)$ gives the inverse logistic function $\sigma(\eta) = \exp(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta})/(1+\exp(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}))$. Polson et al. (2013) propose a data augmentation scheme for logistic regression involving the Pólya-Gamma distribution. A Pólya-Gamma random variable is an infinite sum of gamma random variables, defined more precisely below.

Definition 2.1 (Polson et al. (2013)). *A random variable X has a Pólya-Gamma distribution $b > 0$ and $c \in \mathbb{R}$, denoted $X \sim PG(b, c)$ if*

$$X \overset{d}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2 + c^2/(4\pi^2)},$$

where the $g_k \sim Gamma(b, 1)$ are independent gamma random variables with shape parameter b and rate parameter 1, and $\overset{d}{=}$ indicates equality in distribution. Let $\boldsymbol{X} \sim PG(b, c)$. Let $\cosh(x) = (\exp(x) + \exp(-x))/2$. The probability density function of X can be written as

$$p(x|b, c) = \cosh^b(c/2) \frac{2^{b-1}}{\Gamma(b)} \sum_{n=0}^{\infty} (-1)^n \frac{\Gamma(n + b)}{\Gamma(n + 1)} \frac{(2n + b)}{\sqrt{2\pi x^3}} \exp\left(-\frac{(2n + b)^2}{8x} - \frac{c^2}{2} x\right).$$

For each binary observation y_i we introduce a latent Pólya-Gamma random variable z_i with parameters $b = 1, c = 0$. Polson et al. show that the likelihood contribution of observation i can be written as an integral over the latent z_i,

$$\begin{aligned} p(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}) &= \frac{[\exp(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta})]^{y_i}}{1 + \exp(\boldsymbol{x}_i^\mathsf{t}\boldsymbol{\beta})} \\ &= 2^{-1} \exp(\kappa_i \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}) \int_0^\infty \exp(-z_i(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta})^2/2) p(z_i|1, 0)\, dz_i, \end{aligned}$$

where $\kappa_i = y_i - 1/2$ and $p(z_i|1, 0)$ is the density of a Pólya-Gamma random variable. The complete data likelihood can be written as a quadratic function of $\boldsymbol{\beta}$. Let $\boldsymbol{X}$ give the $n \times p$ design matrix where row i is given by $\boldsymbol{x}_i^\mathsf{T}$. Up to arbitrary constants, the complete data likelihood is equal to

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\beta}) &= p(\boldsymbol{z}) \prod_{i=1}^{n} \exp(\kappa_i \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} - z_i(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta})^2/2) \\ &= p(\boldsymbol{z}) \prod_{i=1}^{n} \exp(\kappa_i^2/(2z_i)) \exp\left(-\frac{z_i}{2}(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} - \kappa_i/z_i)^2\right) \\ &= h(\boldsymbol{z}) p(\boldsymbol{z}) \exp\left(\sum_{i=1}^{n} -\frac{z_i}{2}(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} - \kappa_i/z_i)^2\right) \\ &= h(\boldsymbol{z}) p(\boldsymbol{z}) \exp\left(-\frac{1}{2}(\boldsymbol{y}' - \boldsymbol{X}\boldsymbol{\beta})^\mathsf{T}\Omega(\boldsymbol{y}' - \boldsymbol{X}\boldsymbol{\beta})\right), \end{aligned}$$

where $\boldsymbol{y}' = (\kappa_1/z_1, \ldots, \kappa_n/z_n)$ and $\Omega = \text{diag}(z_1, \ldots, z_n)$ and $h(\boldsymbol{y}, \boldsymbol{z})$ is a function of the augmented dataset that is not dependent on $\boldsymbol{\beta}$. This resembles a regression likelihood with working responses $\boldsymbol{y}'$, design matrix $\boldsymbol{X}$, coefficients $\boldsymbol{\beta}$, and diagonal covariance matrix Ω^{-1}. With the connection to linear regression, it is simple to see that a conditionally conjugate prior on $\boldsymbol{\beta}$ is a multivariate normal distribution. The enriched parametrisation (2.51) is cumbersome to work with, we instead use the more typical parametrisation in terms of the prior mean $\boldsymbol{m}_0$ and prior covariance matrix $\boldsymbol{V}_0$. A two-block Gibbs sampler can be used to sample from the joint posterior of the coefficients $\boldsymbol{\beta}$ and the latent variables $\boldsymbol{z}$. The full conditional on $\boldsymbol{\beta}$ is

$$p(\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{y}) = N(\boldsymbol{m}, \boldsymbol{V}), \tag{2.71}$$

where

$$\begin{aligned} \boldsymbol{V} &= (\boldsymbol{X}^\mathsf{T}\boldsymbol{\Omega}\boldsymbol{X} + \boldsymbol{V}_0^{-1})^{-1} \\ \boldsymbol{m} &= \boldsymbol{V}(\boldsymbol{X}^\mathsf{T}\boldsymbol{\kappa} + \boldsymbol{V}_0^{-1}\boldsymbol{m}_0). \end{aligned}$$

Recall $\boldsymbol{\kappa} = (y_1 - 1/2, \ldots, y_n - 1/2)$ and $\boldsymbol{\Omega}$ is a diagonal matrix where $\Omega_{ii} = z_i$. The latent variables $\boldsymbol{z}$ are conditionally independent given $\boldsymbol{y}$ and $\boldsymbol{\theta}$. The full conditional for the latent z_i is

$$p(z_i|\boldsymbol{\beta}, \boldsymbol{y}) = PG(1, \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}), \tag{2.72}$$

for $i = 1, \ldots, n$, and $PG(b, c)$ denotes a Pólya-Gamma distribution with parameters (b, c). The Pólya-Gamma distribution is not in the exponential family, but still has a useful conditional conjugacy property for the logistic regression model.

From section 2.5.5 the subposterior distributions can also be targeted using Gibbs sampling. The subprior is a $N(\boldsymbol{m}_0, s\boldsymbol{V}_0)$ distribution. The fractionation scales the covariance matrix by a factor of s. To sample from $\widetilde{p}(\boldsymbol{\beta}, \boldsymbol{z}_i|\boldsymbol{y}_i)$ we iterate sampling from $\widetilde{p}(\boldsymbol{\beta}|\boldsymbol{z}_i, \boldsymbol{y}_i)$ and $\widetilde{p}(\boldsymbol{z}_i|\boldsymbol{\beta}, \boldsymbol{y}_i)$. Let $\boldsymbol{x}_{ij}$ give the vector of covariates for the jth observation in subset i for $j \in \{1, \ldots, n_i\}$ and $i \in \{1, \ldots, s\}$. Let $\boldsymbol{X}_{(i)}$ give the $n_i \times p$ matrix of covariates in subset i for $i = 1, \ldots, s$. The subposterior full conditional on $\boldsymbol{\beta}$ is

$$\widetilde{p}(\boldsymbol{\beta}|\boldsymbol{z}_i, \boldsymbol{y}_i) = N(\boldsymbol{m}, \boldsymbol{V}), \tag{2.73}$$

where

$$\begin{aligned}\boldsymbol{V} &= (\boldsymbol{X}_{(i)}^{\mathsf{T}}\boldsymbol{\Omega}\boldsymbol{X}_{(i)} + s^{-1}\boldsymbol{V}_0^{-1})^{-1} \\ \boldsymbol{m} &= \boldsymbol{V}(\boldsymbol{X}_{(i)}^{\mathsf{T}}\boldsymbol{\kappa} + s^{-1}\boldsymbol{V}_0^{-1}\boldsymbol{m}_0).\end{aligned}$$

Here $\boldsymbol{\kappa} = (y_{i1} - 1/2, \ldots, y_{in_i} - 1/2)$ and $\boldsymbol{\Omega} = \text{diag}(\boldsymbol{z}_i)$. The latent variables $\boldsymbol{z}_i$ are again conditionally independent given $\boldsymbol{y}_i$ and $\boldsymbol{\beta}$. The full conditional for the latent z_{ij} is

$$\widetilde{p}(z_{ij}|\boldsymbol{\beta}, \boldsymbol{y}_i) = PG(1, \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta}), \tag{2.74}$$

for $j = 1, \ldots, n_i$. Suppose we run B sweeps of the Gibbs sampler in each subset analysis during the apply step. As dictated by Algorithm 2.2, during the apply step we save the parameters of the full conditional $\widetilde{p}(\boldsymbol{\beta}|\boldsymbol{y}_i, \boldsymbol{z}_i)$ at each iteration. Let $\boldsymbol{m}_i^{[b]}$ represent the conditional mean of $\boldsymbol{\beta}$ and let $\boldsymbol{V}_i^{[b]}$ represent the conditional variance of $\boldsymbol{\beta}$ at iteration b in the ith subposterior analysis for $i = 1, \ldots, s$ and $b = 1, \ldots, B$. Likewise, let $\boldsymbol{z}_i^{[b]}$ denote the sampled latent variables at iteration b in the ith subset analysis for $i = 1, \ldots, s$ and $b = 1, \ldots, B$.

Each worker can compute the subposterior evidence using Chib's method. For any ordinate $\boldsymbol{\beta}^*$ subsposterior evidence satisfies the identity

$$\log \widetilde{p}(\boldsymbol{y}_i) = \log p(\boldsymbol{y}_i|\boldsymbol{\beta}^*) + \log \widetilde{p}(\boldsymbol{\beta}^*) - \log \widetilde{p}(\boldsymbol{\beta}^*|\boldsymbol{y}_i).$$

A simulation consistent estimator is therefore

$$\begin{aligned}\widehat{\log}\ \widetilde{p}(\boldsymbol{y}_i) &= \log p(\boldsymbol{y}_i|\boldsymbol{\beta}^*) + \log \widetilde{p}(\boldsymbol{\beta}^*) - \log\left(B^{-1}\sum_{b=1}^{B} p(\boldsymbol{\beta}^*|\boldsymbol{y}_i, \boldsymbol{z}_i^{[b]})\right) \\ &= \log p(\boldsymbol{y}_i|\boldsymbol{\beta}^*) + \log \widetilde{p}(\boldsymbol{\beta}^*) - \log\left(B^{-1}\sum_{b=1}^{B} N(\boldsymbol{\beta}^*; \boldsymbol{m}_i^{[b]}, \boldsymbol{V}_i^{[b]})\right).\end{aligned} \tag{2.75}$$

We now turn to the combine step. The subposterior integral has the representation

$$I_{\text{sub}} = \mathbb{E}_{\widetilde{p}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)}\left[\int_\Omega \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\beta}|\boldsymbol{y}_i, \boldsymbol{z}_i)\ d\boldsymbol{\beta}\right].$$

From the results in section 2.5.6 it is possible to obtain a closed form expression for the integral over the augmented subposterior distributions given the conjugate structure of the model. Here the conditional subposterior is multivariate normal. Lemma 2.1 gives a result for the integral over a product of multivariate Gaussians.

Lemma 2.1. *Let $p(\boldsymbol{x}; \boldsymbol{\mu}_i, \Sigma_i)$ denote a multivariate Gaussian pdf with mean vector $\boldsymbol{\mu}_i$ and covariance matrix Σ_i for $i = 1, \ldots, s$. For $i = 1, \ldots, s$ define*

$$\begin{aligned}\boldsymbol{\eta}_i &= \Sigma_i^{-1}\boldsymbol{\mu}_i, \\ \Lambda_i &= \Sigma_i^{-1}, \\ \xi_i &= -\frac{1}{2}\left(d\log 2\pi - \log|\Lambda_i| + \boldsymbol{\eta}_i^{\mathsf{T}}\Lambda_i\boldsymbol{\eta}_i\right).\end{aligned}$$

Similarly, define

$$\begin{aligned}\boldsymbol{\eta} &= \textstyle\sum_{i=1}^{s} \boldsymbol{\eta}_i, \\ \Lambda &= \textstyle\sum_{i=1}^{s} \Lambda_i \\ \xi &= -\frac{1}{2}\left(d \log 2\pi - \log|\Lambda| + \boldsymbol{\eta}^{\mathsf{T}} \Lambda \boldsymbol{\eta}\right).\end{aligned}$$

The integral over the product of the s density functions has the closed form solution

$$\int \prod_{i=1}^{s} p(\boldsymbol{x}; \boldsymbol{\mu}_i, \Sigma_i)\ d\boldsymbol{x} = \exp\left[(\textstyle\sum_{i=1}^{s} \xi_i) - \xi\right].$$

The proof uses the fact that the product of multivariate Gaussian density functions is proportional to another multivariate Gaussian density function. This is a special case of the general result in section 2.5.6. For convenience define the function $f(\boldsymbol{m}_1, \ldots, \boldsymbol{m}_s, \Sigma_1, \ldots, \Sigma_s)$ as

$$f(\boldsymbol{m}_1, \ldots, \boldsymbol{m}_s, \boldsymbol{V}_1, \ldots, \boldsymbol{V}_s) = \int \prod_{i=1}^{s} N(\boldsymbol{x}; \boldsymbol{m}_i, \boldsymbol{V}_i)\ d\boldsymbol{x}.$$

The function $f(\cdot)$ can be numerically evaluated using the result in Lemma 2.1. The Monte-Carlo estimator of the subposterior integral is then

$$\begin{aligned}\widehat{I}_{\text{sub}} &= B^{-1} \sum_{b=1}^{B} \int_{\Omega} \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\beta}|\boldsymbol{y}_i, \boldsymbol{z}_i^{[b]})\ d\boldsymbol{\beta} \\ &= B^{-1} \sum_{b=1}^{B} f(\boldsymbol{m}_1^{[b]}, \ldots, \boldsymbol{m}_s^{[b]}, \boldsymbol{V}_1^{[b]}, \ldots, \boldsymbol{V}_s^{[b]}) \end{aligned} \tag{2.76}$$

The subprior normalising constant is

$$\alpha = s^{p/2} (\sqrt{(2\pi)^p |\boldsymbol{V}_0|})^{1-1/s}.$$

Combining (2.75) and (2.76) the full dataset model evidence is estimated as

$$\widehat{\log}\ p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\boldsymbol{\beta}) = \left(\sum_{i=1}^{s} \widehat{\log}\ \widetilde{p}(\boldsymbol{y}_i)\right) + s \log \alpha + \log \widehat{I}_{\text{sub}}.$$

2.7 Data application

2.7.1 Flights dataset

We considered a dataset on flights departing New York city, available in the R package `nycflights13` (Wickham, 2014). There are $n = 327,346$ observations. We dichotomised the arrival delay variable (original units in minutes) to obtain a binary outcome. We labelled flights as late if the arrival delay was greater than zero, and on time if the arrival delay was less than or equal to zero. Unsurprisingly, there is a clear statistical association between late arrival at the destination and the departure delay leaving New York. There are data on 16 different carriers (airlines).

Figure 2.9 (a) plots the number of flights against departure delay. Panel (b) plots the proportion of flights that were late against departure delay in minutes. In (a) we see that the distribution of departure delay time has a mode slightly above zero. The majority of flights depart late. In (b) we see a logistic relationship between departure delay (minutes) and the empirical probability of late arrival at the destination. This suggests a logistic regression model is appropriate.

Let $y_i \in \{0, 1\}$ denote the response, where $y_i = 1$ if flight i was late and is zero otherwise. We considered two regression models for the response,

$$\mathcal{M}_1 = \text{intercept} + \text{delay} \tag{2.77}$$

$$\mathcal{M}_2 = \text{intercept} : \text{carrier} + \text{delay} : \text{carrier}. \tag{2.78}$$

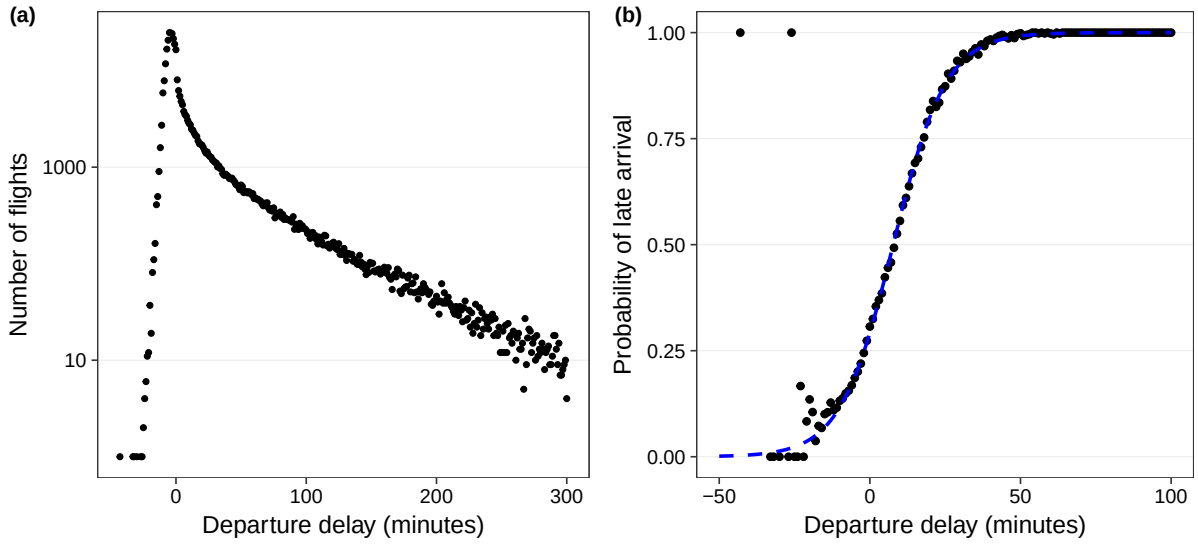


Figure 2.9: Flights dataset ($n = 327,346$) and fitted logistic model. (a) shows the number of flights against departure delay. (b) Probability of late arrival given departure delay. Points are observed data. The blue solid line is the posterior mean from the logistic regression model. In (a) we see that the distribution of departure delay time has a mode slightly above zero. The majority of flights depart late. In (b) we see a logistic relationship between departure delay (minutes) and the empirical probability of late arrival at the destination.

Model 1 is a pooled model where all carriers are modelled identically. Model 2 is a completely unpooled model where each carrier is given a unique intercept and slope. We used independent $N(0,1)$ priors on the coefficients. In panel (b) of Figure 2.9 we plot the posterior predictive mean fit of model 1 as a dashed line. The posterior predictive mean for a new response given covariates $\boldsymbol{x}_{\text{new}}$ is obtained by integrating over the posterior distribution of the coefficients

$$\mathbb{E}[y_{\text{new}}|\mathcal{M}] = \int p(y_{\text{new}} = 1|x_{\text{new}}, \boldsymbol{\beta}, \mathcal{M}) p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y}, \mathcal{M}). \tag{2.79}$$

The fit of $\mathcal{M}_1$ to the observed data appears to be extremely good. There is some deviation from the theoretical proportion for very early flights, say with a departure delay of less than -20 minutes. However, we have very few data points in this range (refer back to panel (a)) so it seems reasonable to see this level of noise in (b).

Figure 2.10 plots probability of late arrival against departure delay for each carrier (airline) in the dataset. The blue dashed line is the fitted mean from the pooled model ($\mathcal{M}_1$). The goodness of fit varies over airlines. There is noticeable deficiency in the pooled model in the results for the carriers FL and MQ. The graphical diagnositic in panel (b) of Figure 2.9 did obscure some deficiencies in the model. When averaging over carriers, model 1 looks very good. However when looking at the quality of the fit for individual carriers we see some systematic problems. Model 2 allows for more flexibility. The posterior mean fit for model 2 is plotted as a solid red line. The fit is noticeably better, but requires the addition of 30 parameters. We computed the integrated likelihood using the divide and conquer strategy with $s = 5$ subsets. Datasets were split uniformly at random. The log Bayes factor in favour of model 2 is 2353. There is very strong support for modelling carriers separately.

2.7.2 Pima Indians dataset

We also analyse the Pima Indians diabetes dataset (Venables and Ripley, 2002). This is not a Big Data application as $n = 532$. The idea is to illustrate how the initial data split affects the Monte Carlo variance of $\widehat{I}_{\text{sub}}$ on a benchmark dataset. We take diabetes status as the response, and glucose and body mass index as the covariates. Let y_j denote a binary response where y_j is equal to subject j has diabetes and is zero otherwise for $j = 1, \ldots, n$. Let $x_{j,1}$ represent glucose level and $x_{j,2}$ represent body

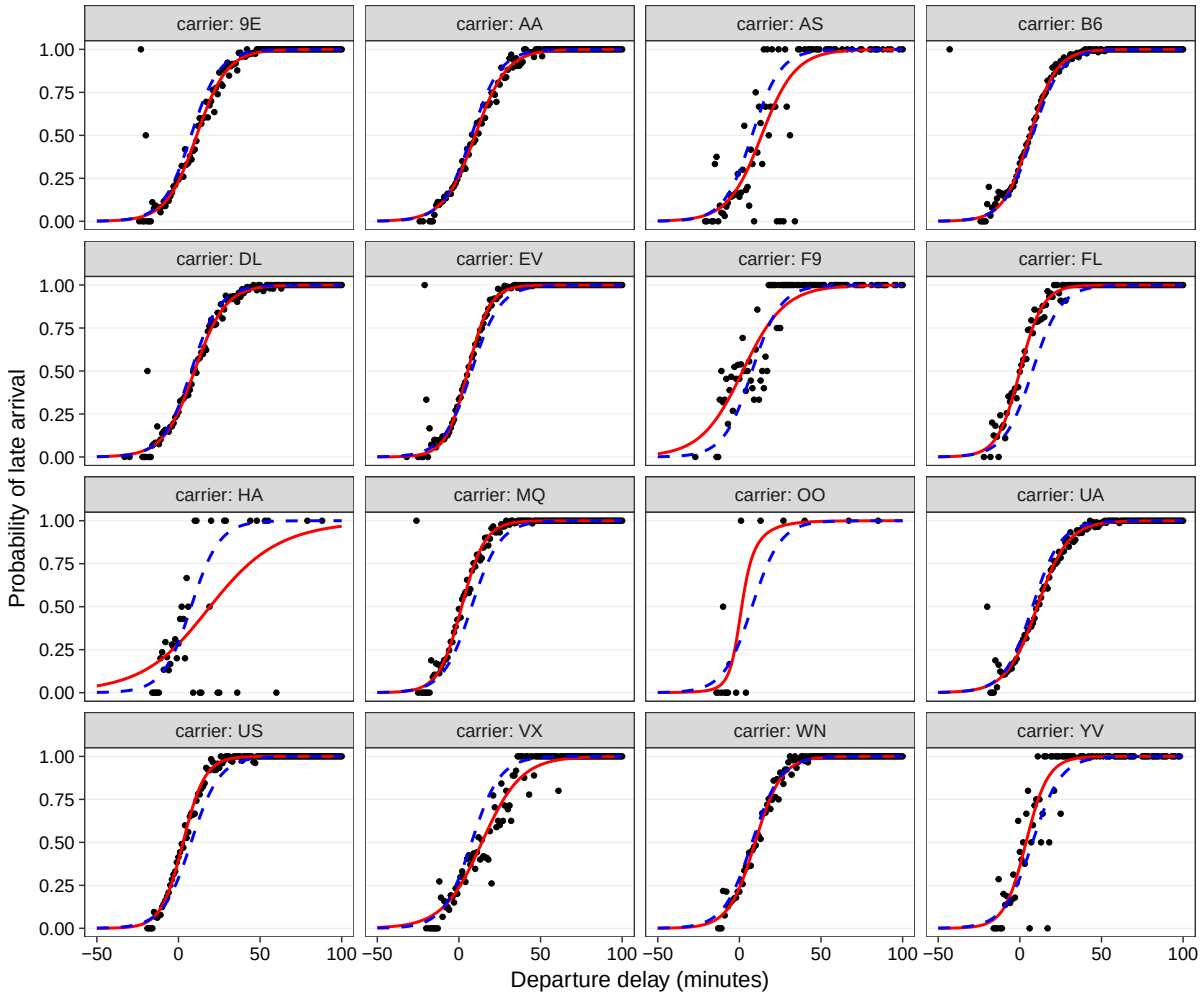


Figure 2.10: Comparison of pooled logistic regression model to unpooled logistic regression model on the flights dataset. The blue dashed line shows the posterior mean fit from the pooled model. The red solid line shows the posterior mean fit from the unpooled model. There appears to be heterogeneity across carriers. The unpooled model gives a better individual fit to each carrier.

mass index for subject j. Given covariates $\boldsymbol{x}_j = (x_{j,1}, x_{j,2})^\mathsf{T}$, y_j is modelled as Bernoulli$(\sigma(\eta_j))$ where $\eta_j = \beta_0 + x_{j,1}\beta_1 + x_{j,2}\beta_2$ and $\sigma(\eta_j) = 1/(1 + \exp(-\eta_j))$.

We compare two partitions of the full dataset $\boldsymbol{y} = (y_1, \ldots, y_n)$ into $s = 2$ subsets $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. The first split is made uniformly at random and is shown in Figure 2.12. The second split oversamples cases by a factor of 10 in the first subset. The biased split of the dataset is shown in Figure 2.13. In each panel we plot a decision boundary from the respective analysis. Let $\boldsymbol{x} = (x_1, x_2)^\mathsf{T}$ represent a vector of covariates. The displayed decision boundary is given by the hyperplane $\boldsymbol{x}^\mathsf{T}\boldsymbol{w} + \alpha = 0$, where $\boldsymbol{w}$ and α are determined using the posterior distribution in the full dataset results and the subposterior distributions in the subset results. In the full dataset analysis we obtain the slope coefficients $\boldsymbol{w} = \mathbb{E}[(\beta_1, \beta_2)^\mathsf{T}|\boldsymbol{y}]$ and the intercept as $\alpha = \mathbb{E}[\beta_0|\boldsymbol{y}]$ where the expectation is over the posterior distribution $p(\beta_0, \beta_1, \beta_2|\boldsymbol{y})$. For the subset results we set $\boldsymbol{w} = \mathbb{E}[(\beta_1, \beta_2)^\mathsf{T}|\boldsymbol{y}_i]$ and the intercept as $\alpha = \mathbb{E}[\beta_0|\boldsymbol{y}]$ where the expectation is over the subposterior distribution $\widetilde{p}(\beta_0, \beta_1, \beta_2|\boldsymbol{y}_i)$ for $i = 1, 2$. The subset decision boundaries are influenced by the subset specific case proportion in $\boldsymbol{y}_i$. In Figure 2.12 the decision boundaries in the subset analyses are similar to the decision boundary in the full dataset analyses. The subset results are in line with the full dataset results. In constrast, in Figure 2.13 we see that the subset decision boundaries are different to the decision boundary from the full dataset analysis. In the subset 2 results the decision boundary is out of the range of the panel. The biased split has caused the subset analyses to not be consistent with the full dataset analysis. This will have a flow on impact to the subposterior distributions on the latent

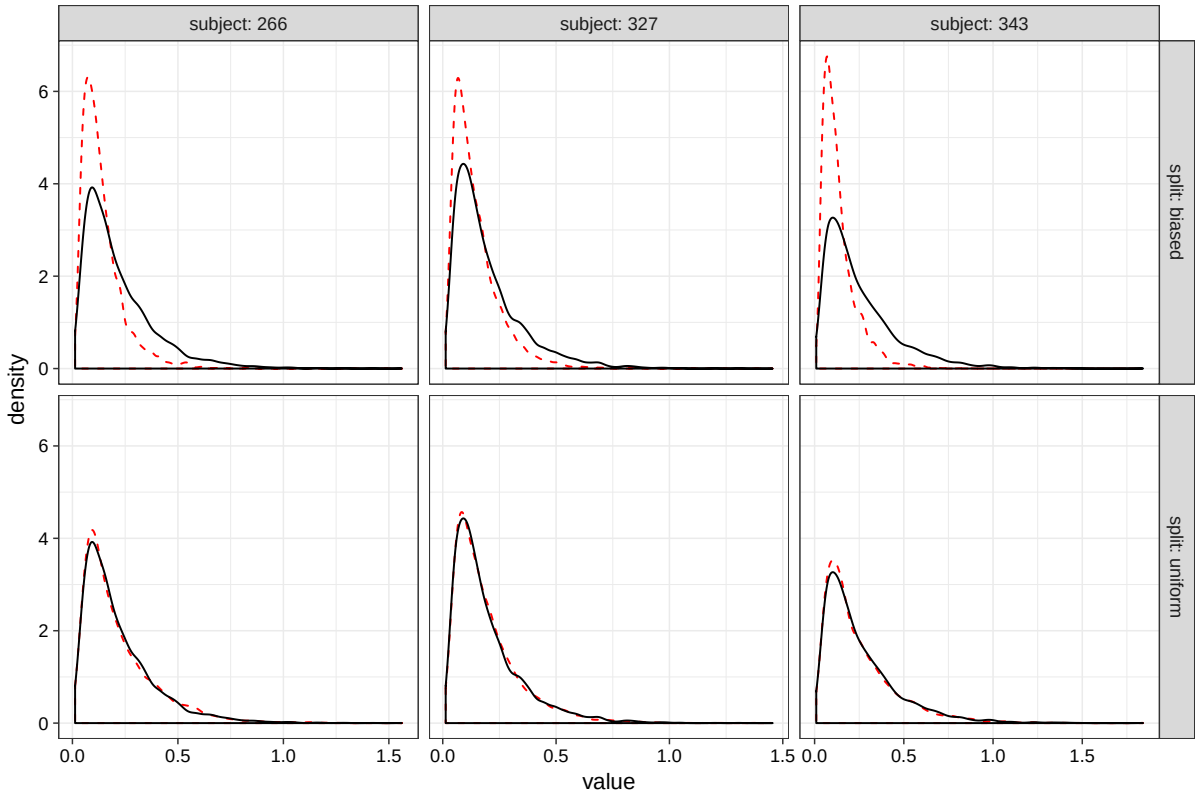


Figure 2.11: Comparison of subposterior and target posterior distributions on the latent variables for the Pima Indians dataset. The solid black line represents the target posterior distribution on the latent variable $p(z_{ij}|\boldsymbol{X}, \boldsymbol{y})$ and the dashed red line gives the subposterior distribution on the latent variable $\widetilde{p}(z_{ij}|\boldsymbol{X}_{(i)}, \boldsymbol{y}_i)$ for the subjects listed in Table 2.6. The disparity between the subposterior and the target posterior distributions is greater under the biased split than under the uniform split. This disparity is directly related to the variance of the subposterior integral estimator (2.70)

variables.

We ran Algorithm 2 with $B = 1000$ for both partitions one hundred times. We then compared the variance of $\log \widehat{I}_{\text{sub}}$. The variance of $\log \widehat{I}_{\text{sub}}$ is a function of the variance of the importance weights

$$\frac{p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s)}{\widetilde{p}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s)}. \tag{2.80}$$

The subposterior distribution $\widetilde{p}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s)$ is expected to be very different to the target distribution $p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s)$ under the biased split. The subposterior distribution $\widetilde{p}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s)$ should be more similar to the target distribution $p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s)$ under the uniform split. The biased split affects the estimate of the intercept parameter β_0 in each subset. This is turn affects the subposterior distribution over the latent variables. Figure 2.11 compares the subposterior distributions on the latent variables to the target distribution for each partition. Results are shown for three observations, described in Table 2.6. In Figure 2.11, the solid line gives the target posterior distribution $p(z_{ij}|\boldsymbol{y}_1, \boldsymbol{y}_2)$ for the three subjects of interest. The dashed line gives the subposterior distribution $\widetilde{p}(z_{ij}|\boldsymbol{y}_1, \boldsymbol{y}_2)$ for each subject. As predicted, the biased split causes $\widetilde{p}(z_{ij}|\boldsymbol{y}_1, \boldsymbol{y}_2)$ to be very different from $p(z_{ij}|\boldsymbol{y}_1, \boldsymbol{y}_2)$. The uniform split results in $\widetilde{p}(z_{ij}|\boldsymbol{y}_1, \boldsymbol{y}_2)$ being very similar to $p(z_{ij}|\boldsymbol{y}_1, \boldsymbol{y}_2)$. As expected, the Monte Carlo variance of $\log \widehat{I}_{\text{sub}}$ is very different under each split. Figure 2.14 shows boxplots of the estimates of $\log \widehat{I}_{\text{sub}}$ obtained in the combine step over the one hundred replications of Algorithm 2.2. The results under the biased split are more variable than the results under the uniform split. The standard deviation under the biased split is 3.04. The standard deviation under the uniform split is 0.032. The initial partition in the split step has a large influence on the Monte Carlo error in the combine step.

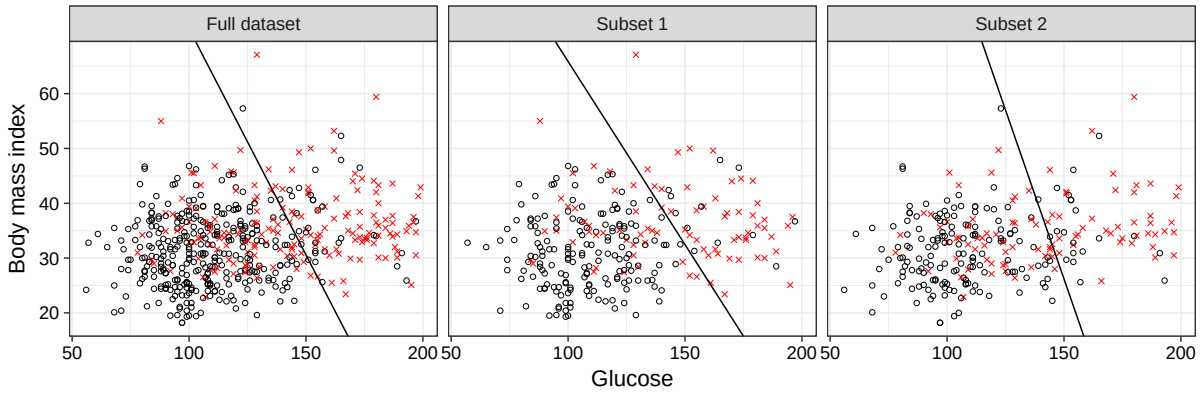


Figure 2.12: Uniform split of the Pima Indians dataset. Solid line gives the decision boundary using the (sub)posterior mean of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{\mathsf{T}}$ in each dataset. Red crosses denote cases and black circles denote controls. The decision boundaries are similar across each analysis. The consistency in the results suggests the evidence synthesis in the combine step will be relatively congenial.

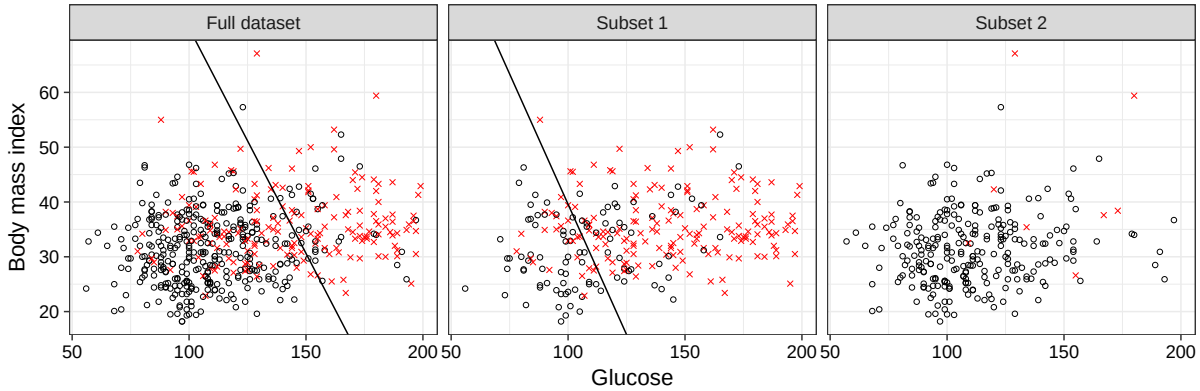


Figure 2.13: Biased split of the Pima Indians dataset. Solid line gives the decision boundary using the (sub)posterior mean of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{\mathsf{T}}$ in each dataset. Red crosses denote cases, and black circles denote controls. The decision boundaries are dissimilar over the analyses. The deviation in the results suggests the evidence synthesis in the combine step will be comparatively challenging.

subject	type	glu	bmi
266	No	106	30.50
327	No	78	36.90
343	No	157	25.60

Table 2.6: Raw data for subjects shown in Figure 2.11.

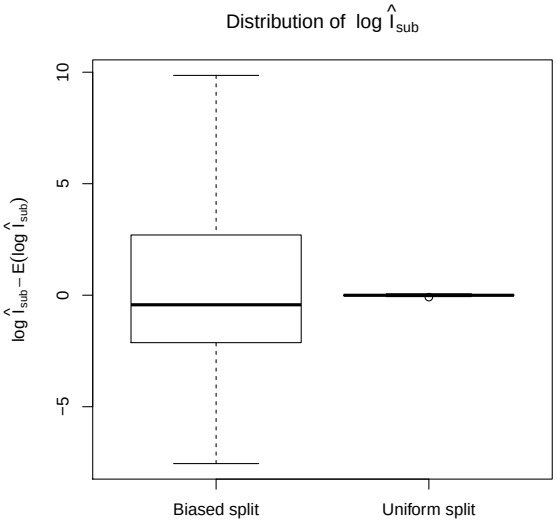


Figure 2.14: Distribution of $\log \widehat{I}_{\text{sub}}$ under the biased split and the uniform split on the Pima Indians dataset. There is greater variance under the biased split compared to the uniform split. The initial partition in the split step influences the Monte Carlo variance in the combine step.

2.8 Conclusion

Parallel processing is a compelling computational pathway for scalable Bayesian inference. Distributed computing platforms facilitate a divide and conquer approach where the Big Data analysis problem can be broken down into a series of smaller conventional analyses. This idea has been explored for posterior simulation. We have investigated divide and conquer Bayesian model selection and found that model selection requires different theory and methods.

We have considered two different algorithms that fit the mould prescribed by Figure 2.1. We initially considered a general approach that is applicable to any any parametric model. MCMC can be used in the apply stage, with the output then consisting of the posterior draws. The combine step can then be approached as a Bayesian hypothesis testing problem. The posterior output can then be used to estimate a Savage-Dickey densty ratio that checks for the global suitability of the model over all subsets. The curse of dimensionality yields a major hurdle for this algorithm in the combine step. The dimension of the density estimation task increases with the number of subsets s, thus limiting the scalability of Algorithm 2.1.

If the original model can be augmented with suitable latent variables, it is possible to obtain a practical algorithm. We propose an embarrassingly parallel algorithm for computing the model evidence that makes use of Gibbs sampling in the apply stage to ease the difficulty of the combine step. The subset Gibbs runs in the apply phase can be undertaken using simple modifications of standard update formulae. The combine step can be easily by carried out aggregating the Gibbs output. Interestingly, the initial split of the dataset strongly influences the Monte Carlo variance in the combine step. If the initial split it very poor, the results in the combine stage will be useless. As such it may be worth exploring adaptive procedures that optimise the data split after a preliminary Gibbs run.

The split-apply-combine methodology using data augmentation may also be of use for posterior sampling. Using the divide and conquer procedure, we generate samples from the subposterior distribution of the latent variables $\widetilde{p}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)$. If we can determine the importance ratio $\widetilde{p}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)$ up to a constant of proportionality, we can use $\widetilde{p}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)$ in a self-normalised importance sampler to target the true posterior. Equation (2.69) showed that the subposterior

integral is proportional to the required importance sampling weights:

$$\int \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{z}_i)\ d\boldsymbol{\theta} = I_{\text{sub}} \times \frac{p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)}{\widetilde{p}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)} \propto \frac{p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)}{\widetilde{p}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)}.$$

Given that we have a closed form expression for the augmented subposterior integral

$$\int_{\Omega} \prod_{i=1}^{s} \widetilde{p}(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{z}_i)\ d\boldsymbol{\theta} = \frac{\prod_{i=1}^{s} c(\nu_0/s + n_i, \boldsymbol{\phi}_0/s + t(\boldsymbol{y}_i, \boldsymbol{z}_i), \boldsymbol{\omega}_0/s)}{c(\nu_0 + n, \boldsymbol{\phi}_0 + \sum_{i=1}^{s} t(\boldsymbol{y}_i, \boldsymbol{z}_i), \boldsymbol{\omega}_0)},$$

we can compute the required importance sampling weights. We then can use self-normalised importance sampling in the combine stage to target the full dataset posterior $p(\boldsymbol{\theta}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s)$. Existing divide and conquer procedures have been criticised for not targeting the exact target posterior (Bardenet et al., 2017). Using the Gibbs based methodology here we can target the exact posterior providing that the model has some conditionally conjugate structure. This is a promising future research direction.

Calculation of the model evidence is an important task, especially so when given masses of data with multiple plausible models. Distributed computing appears to be useful in the pursuit of this goal.

3

Bounding the model evidence using the subsampled sandwich estimator

Summary

We investigate computational methods for estimation of the integrated likelihood on large n datasets. Existing importance sampling techniques can be impractical due to the need for full likelihood evaluations. We develop a novel strategy for estimating the log model evidence using subsampling. We propose to pair a variational lower bound with an upper bound formed using a maximum entropy argument. The enclosing bounds are asymptotically tight as n grows, and permit a reduction in the computational budget. The bounds can be estimated in a computationally efficient manner using subsampling and control variates. The sandwich approach provides more definitive error measures than other methods for approximating the integrated likelihood. We demonstrate the methods on a large logistic regression dataset.

3.1 Introduction

The computational expense of likelihood evaluations is the root cause of many challenges in Bayesian computation for Big Data. Standard iterative algorithms can struggle with large datasets as the $O(n)$ cost per loop stalls the procedure from a practical point of view. A generic goal is to reduce computational cost by allowing for access of only a small minibatch of data per iteration. Most existing algorithms for estimating the integrated likelihood suffer from the likelihood scaling issue, becoming impractical for huge n. We develop a novel strategy for estimating the model evidence on tall datasets, based on sandwiching the log integrated likelihood between upper and lower bounds. Squeezing the log integrated likelihood between encasing bounds is a departure from traditional importance sampling methods that deliver point estimates of the model evidence. Shifting to interval estimation is useful as it eases the integration of subsampling into the algorithm. We find it is necessary to use control variates to obtain a scalable algorithm, a common finding in the Big Data literature.

Subsampling has been examined as a technique to reduce the computational cost of Bayesian inference in the fixed model setting, where the primary goal is to generate samples from the posterior distribution (Maclaurin and Adams, 2014; Quiroz et al., 2018). Given a likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$ and prior $p(\boldsymbol{\theta})$, suppose we wish to sample from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Markov chain Monte Carlo (MCMC) has proven to be a versatile computational engine for applied Bayesian inference, facilitating sampling from the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ in complex models. As a simple example, consider the Metropolis-Hastings algorithm, described in Algorithm 3.1. The calculation of the acceptance ratio $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}')$ in line 6 requires the full likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$. The $O(n)$ cost per iteration can render the algorithm impractically slow for Big Data applications. There is a body of work exploring the integration of subsampling into the pseudo-marginal MCMC framework of Andrieu et al. (2009) to address this likelihood cost. The innovation behind the pseudo-marginal idea is that if the likelihood calculations in the acceptance ratio are replaced

Algorithm 3.1 Metropolis-Hastings for simulating from the posterior distribution

1: Input $\boldsymbol{\theta}^{[0]}$
2: **for** $b = 1, \ldots, B$ **do**
3: $\quad \boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{[b-1]}$
4: $\quad \boldsymbol{\theta}' \sim g(\cdot|\boldsymbol{\theta})$ ▷ Sample new proposal from g
5:
6: $\quad \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') \leftarrow \frac{p(\boldsymbol{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})} \times \frac{g(\boldsymbol{\theta}|\boldsymbol{\theta}')}{g(\boldsymbol{\theta}'|\boldsymbol{\theta})}$ ▷ acceptance ratio calculation
7:
8: $\quad u \sim \text{Uniform}(0, 1)$
9: $\quad$ **if** $u < \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}')$ **then**
10: $\quad\quad \boldsymbol{\theta}^{[b]} \leftarrow \boldsymbol{\theta}'$ ▷ Accept proposal
11: $\quad$ **else**
12: $\quad\quad \boldsymbol{\theta}^{[b]} \leftarrow \boldsymbol{\theta}$ ▷ Reject proposal
13: $\quad$ **end if**
14: **end for**
15: **return** $\left\{\boldsymbol{\theta}^{[1]}, \ldots, \boldsymbol{\theta}^{[B]}\right\}$

with a non-negative unbiased estimator $\widehat{p}(\boldsymbol{y}|\boldsymbol{\theta})$, it is still possible to define a Markov chain that maintains the correct target distribution $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})$. The pseudo-marginal algorithm for tall datasets is described in Algorithm 3.1. The pseudo-marginal algorithm uses an approximate acceptance ratio in line 8. The key point of difference relative to Algorithm 3.1, is that an estimated likelihood now appears in the numerator of the acceptance ratio. An important detail of the pseudo-marginal algorithm is that previous likelihood estimates need to be recycled if the proposed value $\boldsymbol{\theta}'$ is rejected (see line 16). There has been a large amount of effort in designing almost-surely non-negative unbiased estimators of the likelihood that use subsampling (Bardenet et al., 2017; Quiroz et al., 2016). Although the per iteration cost decreases when using a subsampled pseudo-marginal algorithm compared the the standard Metropolis-Hastings implementation, the autocorrelation in the chain increases with the variance of the likelihood estimator. A highly variable likelihood estimator can potentially reduce the effective sample size per unit time relative to the standard Metropolis-Hastings algorithm (Maclaurin and Adams, 2014).

Algorithm 3.2 Pseudo-marginal Metropolis-Hastings using subsampling

1: Input $\boldsymbol{\theta}^{[0]}$, $p(\boldsymbol{y}|\boldsymbol{\theta})^{[0]}$
2: **for** $b = 1, \ldots, B$ **do**
3: $\quad \boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{[b-1]}$
4: $\quad \widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}) \leftarrow \widehat{p}(y|\boldsymbol{\theta})^{[b-1]}$
5: $\quad \boldsymbol{\theta}' \sim g(\cdot|\boldsymbol{\theta})$ ▷ Sample new proposal from g
6: $\quad$ Estimate likelihood $\widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}')$ ▷ Use subsample of full dataset
7:
8: $\quad \widehat{\alpha}(\boldsymbol{\theta}, \boldsymbol{\theta}') \leftarrow \frac{\widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}{\widehat{p}(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})} \times \frac{g(\boldsymbol{\theta}|\boldsymbol{\theta}')}{g(\boldsymbol{\theta}'|\boldsymbol{\theta})}$ ▷ Acceptance ratio calculation
9:
10: $\quad u \sim \text{Uniform}(0, 1)$
11: $\quad$ **if** $u < \widehat{\alpha}(\boldsymbol{\theta}', \boldsymbol{\theta})$ **then**
12: $\quad\quad \boldsymbol{\theta}^{[b]} \leftarrow \boldsymbol{\theta}'$ ▷ Accept proposal
13: $\quad\quad \widehat{p}(\boldsymbol{y}|\boldsymbol{\theta})^{[b]} \leftarrow \widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}')$ ▷ Update likelihood estimate
14: $\quad$ **else**
15: $\quad\quad \boldsymbol{\theta}^{[b]} \leftarrow \boldsymbol{\theta}$ ▷ Reject proposal
16: $\quad\quad \widehat{p}(\boldsymbol{y}|\boldsymbol{\theta})^{[b]} \leftarrow \widehat{p}(\boldsymbol{y}|\boldsymbol{\theta})^{[b-1]}$ ▷ Retain likelihood estimate
17: $\quad$ **end if**
18: **end for**
19: **return** $\left\{\boldsymbol{\theta}^{[1]}, \ldots, \boldsymbol{\theta}^{[B]}\right\}$

Subsampling has also been explored as a technique to reduce the per epoch cost of other stochastic simulation methods. Stochastic gradient Langevin dynamics (Welling and Teh, 2011) bypasses the accept reject step entirely, the iterative algorithm uses subsampled estimates of the gradient of the log

likelihood for computationally cheap individual steps. The Zig-Zag sampler (Bierkens et al., 2016) and the scalable Langevin exact sampler (Pollock et al., 2016) are both based on discretisations of continuous time stochastic processes and use subsampled estimates of the gradient of the log likelihood. Unbiased estimation of the log scale is significantly easier than unbiased estimation of likelihoods, and this is seen as a competitive advantage of non pseudo-marginal based methods (Bardenet et al., 2017).

The mechanics of the aforementioned large n samplers are varied and technical and will not be reviewed here. We will proceed assuming that we have a large dataset and have obtained samples from the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ using a suitable algorithm. The task of interest is to estimate the integrated likelihood $p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\ d\boldsymbol{\theta}$. Given posterior samples, it is common to use importance sampling based methods for estimating the integrated likelihood $p(\boldsymbol{y})$ (Friel and Wyse, 2012). Importance sampling methods for computing the evidence typically require repeated additional full likelihood evaluations, making them unsuited to large n problems. This is a roadblock for carrying out Bayesian model selection on tall datasets. We will argue that it is difficult to adapt existing importance sampling methods to use subsampling effectively. Faster methods for approximating the integrated likelihood include the Laplace approximation, the Laplace-Metropolis estimator, and variational Bayesian methods. Although computationally cheap, it is difficult to bound the approximation error of these faster methods.

The emergence of tall datasets has led to a shift in the procedures used for posterior sampling. Estimation of the model evidence may also require a similar paradigm shift. Historically, estimation of the integrated likelihood is performed using a combination of posterior simulation and asymptotic approximation (DiCiccio et al., 1997). We also use these techniques but with the large n dynamics of the posterior distribution in mind. Underlying our approach is a less commonly used representation of the log model evidence. The log model evidence $\log p(\boldsymbol{y})$ can be written as an information criterion type score involving a goodness of fit term and a penalty term for model complexity. Starting from Bayes' theorem, we easily obtain the so called 'Candidiate's formula' an important identity relating the model evidence to the posterior and prior densities (Besag, 1989). For any ordinate $\boldsymbol{\theta} \in \Omega$ we have the identity :

$$\log p(\boldsymbol{y}) = \log p(\boldsymbol{y}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\boldsymbol{y}) \tag{3.1}$$

$$= \log p(\boldsymbol{y}|\boldsymbol{\theta}) - \log \frac{p(\boldsymbol{\theta}|\boldsymbol{y})}{p(\boldsymbol{\theta})}. \tag{3.2}$$

Taking the Candidate's formula (3.2) one step further, we can integrate both sides over the posterior distribution to obtain the following decomposition of the log model evidence:

$$\int \log p(\boldsymbol{y}) p(\boldsymbol{\theta}|\boldsymbol{y})\ d\boldsymbol{\theta} = \int p(\boldsymbol{\theta}|\boldsymbol{y}) \log p(\boldsymbol{y}|\boldsymbol{\theta})\ d\boldsymbol{\theta} - \int p(\boldsymbol{\theta}|\boldsymbol{y}) \log \frac{p(\boldsymbol{\theta}|\boldsymbol{y})}{p(\boldsymbol{\theta})}\ d\boldsymbol{\theta}.$$

$$\log p(\boldsymbol{y}) = \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{\theta})] - D(p(\boldsymbol{\theta}|\boldsymbol{y}) \parallel p(\boldsymbol{\theta})). \tag{3.3}$$

The identity (3.3) is not an original result, however it seems to have attracted very little attention in the existing literature on Bayesian model selection (Robert, 2007; Gelman et al., 2014; Kass and Raftery, 1995; Gelfand and Dey, 1994; Raftery, 1995). The first term $\mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{\theta})]$ is the expected log likelihood over the posterior distribution. This can be characterised as a Bayesian measure of goodness of fit (Dempster, 1997; Spiegelhalter et al., 2002). The second term $D(p(\boldsymbol{\theta}|\boldsymbol{y}) \parallel p(\boldsymbol{\theta}))$ represents the Kullback-Leibler divergence of the prior from the posterior. This can be seen as a penalty term that captures model complexity and prior regularisation simultaneously. We propose to estimate the goodness of fit term using subsampling, and to bound the penalty term using elementary information theory. The goodness of fit term and the bounds on the penalty can be estimated using a posterior samples and subsampled evaluations of the log likelihood. Overall we obtain a computationally efficient interval estimator of the log model evidence that avoids repeated $O(n)$ likelihood evaluations. Asymptotic arguments suggest that the evidence bounds may be sufficient to rank competing models confidently in the large n regime. The general idea of introducing bounded error in order to drastically cut the computational budget has also

been explored for posterior simulation (Bardenet et al., 2014; Korattikara et al., 2014), and we feel it is natural to follow this avenue for estimation of the model evidence.

The sandwich approach reduces the computational burden of estimating the integrated likelihood while still giving confidence measures on the quality of the estimate. Section 3.2 reviews asymptotic properties of Bayesian model selection and develops the upper and lower bounds on the penalty term. Section 3.3 reviews how subsampling can be used to estimate likelihoods and recaps existing methods for estimating the model evidence. Section 3.4 discusses the utility of our proposed evidence bounds for tall datasets. We test our proposed methodology in a logistic regression data application in Section 3.5.

3.2 Bayesian model selection

Suppose that we have M competing parametric models $\mathcal{M}_1, \ldots, \mathcal{M}_M$. Each model $j = 1, \ldots, M$ has parameter $\boldsymbol{\theta}_j$ which takes values in parameter space Ω_j of dimension p_j. We assume the data $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ consists of n independently and identically distributed observations from some distribution f_0. We take the $\mathcal{M}$-open viewpoint, in that the true generative model f_0 may not be in the set of candidate models $\mathcal{M}_1, \ldots, \mathcal{M}_M$. Given two models $\mathcal{M}_j$ and $\mathcal{M}_k$, let $\log \mathcal{B}_{jk}$ denote the log Bayes factor in favour of model j:

$$\log \mathcal{B}_{jk} = \log p(\boldsymbol{y}_i|\mathcal{M}_j) - \log p(\boldsymbol{y}|\mathcal{M}_k). \tag{3.4}$$

The asymptotic behaviour of $\log \mathcal{B}_{jk}$ is of direct interest given our focus on tall data. For each model $\mathcal{M}_j$ we can define a measure of closeness to the true model f_0 using the Kullback-Leibler divergence. Let $d_{KL}(f_0, g)$ denote the Kullback-Leibler divergence of a candidate density g from the true generate model f_0. Formally,

$$d_{KL}(f_0, g) = \int f_0 \log \frac{f_0}{g}.$$

For each model $\mathcal{M}_j$ we can find a density $p(\boldsymbol{y}_i|\boldsymbol{\theta}_j^*, \mathcal{M}_j)$ such that $\boldsymbol{\theta}_j^*$ satisfies

$$\theta_j^* := \underset{\boldsymbol{\theta}_j \in \Omega_j}{\arg\min}\, d_{KL}(f_0, p(\boldsymbol{y}_i|\boldsymbol{\theta}_j, \mathcal{M}_j)). \tag{3.5}$$

The density $p(\boldsymbol{y}_i|\boldsymbol{\theta}_j^*)$ is the closest density to the true model f_0 in the parametric family associated with model $\mathcal{M}_j$. We can characterise this as the best approximation to the truth attainable under model $\mathcal{M}_j$. Let d_{KL}^j denote the divergence from the truth for model j, that is

$$d_{KL}^j = d_{KL}(f_0, p(\boldsymbol{y}_i|\boldsymbol{\theta}_j^*)). \tag{3.6}$$

Under regularity conditions, as n tends to infinity, the limiting Bayes factors between two competing models j and k will be controlled by d_{KL}^j and d_{KL}^k. The asymptotic behaviour of the posterior distribution on models can be described by the asymptotic behaviour of the set of Bayes factors. In the $\mathcal{M}$-open viewpoint pragmatic consistency of the posterior distribution amounts to ensuring that the posterior distribution concentrates on the best possible model in the candidate set. Bayes factor consistency is described more formally in Definition 3.1.

Definition 3.1 (Pragmatic Bayes Factors Consistency (Walker et al., 2004))**.** *Consider two models $\mathcal{M}_j$ and $\mathcal{M}_k$. Model j has p_j parameters and Model k has p_k parameters. Let d_{KL}^j and d_{KL}^k be defined as in* (3.6)*. We say that the Bayes factors achieve pragmatic consistency if they exhibit the following asymptotic behaviour as $n \to \infty$ for all $j, k \in \{1, \ldots, M\}$.*

(a) Suppose that $d_{KL}^j < d_{KL}^k$. Then we require that $\log \mathcal{B}_{jk} \to \infty$

(a) Suppose that $d_{KL}^j > d_{KL}^k$. Then we require that $\log \mathcal{B}_{jk} \to -\infty$

$2 \log \mathcal{B}_{jk}$	$\mathcal{B}_{jk}$	Strength of evidence
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
>10	>150	Very strong

Table 3.1: Guidelines for the interpretation of Bayes factors (Kass and Raftery, 1995). Given two models $\mathcal{M}_j$ and $\mathcal{M}_k$, the quantity $\mathcal{B}_{jk}$ gives the Bayes factor in favour of model j (recall equation (3.4)).

(a) Suppose that $d^j_{KL} = d^k_{KL}$ *and model* j *is nested in model* k*, so* $p_j < p_k$*. Then we require that* $\log \mathcal{B}_{jk} \to \infty$.

The case of non-nested models with $d^j_{KL} = d^k_{KL}$ is more complicated, and there does not seem to be a general theory for this situation (Casella et al., 2009). We assume that this we do not encounter this problem for the sake of simplicity.

Pragmatic Bayes factor consistency requires that asymptotically all Bayes factors tend to negative infinity or positive infinity. Asymptotically, the posterior distribution of models concentrates tightly on the best model in the candidate set, even under misspecification. Under relatively mild conditions it is possible to show that pragmatic Bayes factor consistency holds for a wide range of problems (Chib and Kuffner, 2016; Chatterjee et al., 2018). Key references for nested models include Schwarz (1978) and Gelfand and Dey (1994). We proceed assuming that the collection of models of interest and data generating process is such that pragmatic Bayes factor consistency holds. The practical interpretation of the consequences of Definition 3.1 is that we expect to see very large Bayes factors when comparing models in the large n regime. Under mild conditions the gaps between the log evidence scores are expected to widen as n grows.

From an inferential point of view, this large sample dynamic needs to be taken into account when interpreting Bayes factors on tall datasets. Kass and Raftery (1995) present a table providing guidelines on the interpretation of Bayes factor, reproduced here as Table 3.1. Bayes factors and p-values do have some commonality in terms of large sample behaviour. The American Statistical Association's statement on p-values drew attention to the fact that it is important to consider the sample size when weighing the evidence provided by a small p-value (Wasserstein and Lazar, 2016). When n is large, practically insignificant deviations from the null hypothesis can lead to very small p-values. This same phenomenon will affect Bayes factors. The log Bayes factors for two competing models are expected to tend to $-\infty$ or ∞ as n increases, even when the two models differ in an practically inconsequential manner. For tall datasets models may become separated by larger factors than what is considered in Table 3.1.

As n increases, exact calculation of the integrated likelihood becomes more computationally demanding. This problem is not insurmountable as the asymptotic behaviour of the posterior distribution on models works in our favour, assuming that we take Definition 3.1 as realistic. As sample sizes increase we expect to see wider margins between the log evidence values within a collection of models. From a decision theoretic point of view, exact calculation of the integrated likelihood becomes less necessary as n increases. We can still rank models confidently given interval estimates of the integrated likelihood if the interval widths are small compared to the true differences in Bayes factors. This large sample behaviour provides some leeway to carry out computationally efficient Bayesian model selection on tall datasets. This is assumption is used to motivate our evidence sandwich approach.

3.2.1 Evidence bounds

An important assumption for our bound to hold is that the parameter $\boldsymbol{\theta}$ has unconstrained support in $\mathbb{R}^d$. It may be necessary to reparameterise the original statistical model in order to meet this condition.

For example, variance components and proportions can be mapped to the real line using the log and logistic transform respectively. We believe this assumption to not be overly restrictive, as a wide range of statistical models can be transformed to an unconstrained parameterisation. The probabilistic programming language Stan (Carpenter et al., 2017) internally transforms all user declared models to an unconstrained parameterisation in order to improve the efficiency of the underlying Hamiltonian Monte Carlo algorithms. The default transformations are listed in the Stan manual (Stan Development Team, 2018). From here on in we assume that $\boldsymbol{\theta}$ has unconstrained support in $\mathbb{R}^d$.

3.2.2 Entropy

The differential entropy represents the amount of uncertainty associated with a continuous random vector. For a continuous random vector $\boldsymbol{X}$ with probability density $p(\boldsymbol{x})$, the differential entropy is defined as

$$\text{differential entropy of } \boldsymbol{X} = -\int p(\boldsymbol{x}) \log p(\boldsymbol{x}) \, d\boldsymbol{x}. \tag{3.7}$$

The differential entropy is not bounded, taking values in $(-\infty, \infty)$. The differential entropy of a multivariate normal random variable is a simple function of the covariance matrix.

Lemma 3.1. *Let $\boldsymbol{X}$ be a d-dimensional multivariate normal random vector $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where Σ is of full rank. The differential entropy of $\boldsymbol{X}$ is given by*

$$\frac{d}{2} + \frac{d}{2} \log(2\pi) + \frac{1}{2} \log(\det(\Sigma)).$$

The result is easily established using properties of quadratic forms. The determinant of the covariance matrix has an interpretation as a scalar measure of the overall variability associated with a multidimensional random vector. The multivariate normal distribution has a maximum entropy property for random vectors in $\mathbb{R}^d$.

Theorem 3.1. *Let $\boldsymbol{X}$ be a random vector in $\mathbb{R}^d$ with mean $\boldsymbol{\mu}$ and full rank covariance matrix Σ. The maximum entropy probability distribution amongst all candidate distributions for $\boldsymbol{X}$ with mean $\boldsymbol{\mu}$ and covariance matrix Σ is the multivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$.*

Proof: Let $q(\boldsymbol{x})$ denote some distribution on $\mathbb{R}^d$ with mean $\boldsymbol{\mu}$ and covariance matrix Σ. Let $p(\boldsymbol{x}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $H(q)$ denote the differential entropy of $q(\boldsymbol{x})$ and let $H(p)$ denote the differential entropy of $p(\boldsymbol{x})$. Let $D(q(\boldsymbol{x}) \parallel p(\boldsymbol{x}))$ denote the Kullback-Leibler divergence of p from q. The Kullback-Leibler divergence is positive so

$$\begin{aligned}
0 &\leq D(q(\boldsymbol{x}) \parallel p(\boldsymbol{x})) \\
&= \int q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} \, d\boldsymbol{x} \\
&= \int q(\boldsymbol{x}) \log q(\boldsymbol{x}) \, d\boldsymbol{x} - \int q(\boldsymbol{x}) \log p(\boldsymbol{x}) \, d\boldsymbol{x} \\
&= -H(q) - \int q(\boldsymbol{x}) \left(-\frac{d}{2} \log(2\pi) - \frac{1}{2} \log(\det(\Sigma)) - \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^\intercal \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right) d\boldsymbol{x} \\
&= -H(q) + \frac{d}{2} \log(2\pi) + \frac{1}{2} \log(\det(\Sigma)) + \frac{d}{2} \\
&= -H(q) + H(p).
\end{aligned}$$

The step in the second last line follows from properties of quadratic forms and the fact that $q(\boldsymbol{x})$ has the same first and second moments as $p(\boldsymbol{x})$. The final line establishes the desired result that $H(p) \geq H(q)$. As $D(q(\boldsymbol{x}) \parallel p(\boldsymbol{x})) = 0$ if and only if $q(\boldsymbol{x})$ is a multivariate normal distribution, $H(q) = H(p)$ if and only if $q(\boldsymbol{x}) = N(\boldsymbol{\mu}, \Sigma)$.

It follows from Theorem 3.1 that given some random d-dimensional vector $\boldsymbol{X}$ with mean $\boldsymbol{\mu}$, full rank covariance matrix Σ and unknown probability density $q(\boldsymbol{x})$, the differential entropy has the upper bound,

$$-\int q(\boldsymbol{x}) \log q(\boldsymbol{x}) \; d\boldsymbol{x} \leq \frac{d}{2} + \frac{d}{2} \log (2\pi) + \frac{1}{2} \log(\det(\Sigma)). \tag{3.8}$$

Equality holds if and only if $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$. We now show how the entropy bound can be used to approximate the model evidence.

3.2.3 Upper bounding the evidence

The penalty term in (3.3) involves a posterior expectation and the negative posterior entropy:

$$D(p(\boldsymbol{\theta}|\boldsymbol{y}) \parallel p(\boldsymbol{\theta})) = p(\boldsymbol{\theta}|\boldsymbol{y}) \log p(\boldsymbol{\theta}|\boldsymbol{y}) - \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{y})} \left[\log p(\boldsymbol{\theta})\right]. \tag{3.9}$$

We can lower bound the penalty term by using the entropy upper bound (3.8). Let $\Sigma_{\boldsymbol{\theta}|\boldsymbol{y}}$ be the posterior covariance matrix of $\boldsymbol{\theta}$. Throughout we make the mild assumption that $\Sigma_{\boldsymbol{\theta}|\boldsymbol{y}}$ is of full rank. The entropy bound (3.8) gives that

$$\int p(\boldsymbol{\theta}|\boldsymbol{y}) \log p(\boldsymbol{\theta}|\boldsymbol{y}) \; d\boldsymbol{\theta} \geq \frac{d}{2} + \frac{d}{2} \log (2\pi) + \frac{1}{2} \log(\det(\Sigma_{\boldsymbol{\theta}|\boldsymbol{y}})).$$

The bound will be very tight when the posterior distribution of $\boldsymbol{\theta}$ is approximately normal. Posterior normality is a plausible assumption when the sample size is large (Van Der Vaart, 1998, Chapter 10). We subsequently obtain the lower bound on the penalty term,

$$D(p(\boldsymbol{\theta}|\boldsymbol{y}) \parallel p(\boldsymbol{\theta})) \geq \frac{d}{2} + \frac{d}{2} \log (2\pi) + \frac{1}{2} \log(\det(\Sigma_{\boldsymbol{\theta}|\boldsymbol{y}})) - \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{y})} \left[\log p(\boldsymbol{\theta})\right]. \tag{3.10}$$

We now upper bound the penalty term.

3.2.4 Lower bounding the evidence

Variational Bayesian inference is a generic family of methods for approximate Bayesian inference (Blei et al., 2017). Variational Bayesian methods introduce an approximate posterior distribution $q(\boldsymbol{\theta})$ where $q(\boldsymbol{\theta})$ is chosen to maximise a lower bound on the log model evidence $\log p(\boldsymbol{y})$. An important identity is that for any distribution $q(\boldsymbol{\theta})$,

$$\log p(\boldsymbol{y}) = \mathbb{E}_{q(\boldsymbol{\theta})} \left[\log p(\boldsymbol{y}|\boldsymbol{\theta})\right] - D(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta})) + D(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta}|\boldsymbol{y})). \tag{3.11}$$

The proof involves a similar line of reasoning that we used to arrive at (3.3). Starting from the 'Candidate's formula' (3.2) we integrate over the arbitrary distribution $q(\boldsymbol{\theta})$.

$$\begin{aligned}
\log p(\boldsymbol{y}) &= \log p(\boldsymbol{y}|\boldsymbol{\theta}) + \log \frac{p(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})} \\
&= \int q(\boldsymbol{\theta}) \left[\log p(\boldsymbol{y}|\boldsymbol{\theta}) + \log \frac{p(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})}\right] \; d\boldsymbol{\theta} \\
&= \mathbb{E}_{q(\boldsymbol{\theta})} \left[\log p(\boldsymbol{y}|\boldsymbol{\theta})\right] + \mathbb{E}_{q(\boldsymbol{\theta})} \left[\log p(\boldsymbol{\theta})\right] + \int q(\boldsymbol{\theta}) \log \frac{1}{p(\boldsymbol{\theta}|\boldsymbol{y})} \; d\boldsymbol{\theta} \\
&= \mathbb{E}_{q(\boldsymbol{\theta})} \left[\log p(\boldsymbol{y}|\boldsymbol{\theta})\right] + \mathbb{E}_{q(\boldsymbol{\theta})} \left[\log p(\boldsymbol{\theta})\right] + \int q(\boldsymbol{\theta}) \log \frac{1}{p(\boldsymbol{\theta}|\boldsymbol{y})} \; d\boldsymbol{\theta} + \\
&\quad \int q(\boldsymbol{\theta}) \log q(\boldsymbol{\theta}) \; d\boldsymbol{\theta} - \int q(\boldsymbol{\theta}) \log q(\boldsymbol{\theta}) \; d\boldsymbol{\theta} \\
&= \mathbb{E}_{q(\boldsymbol{\theta})} \left[\log p(\boldsymbol{y}|\boldsymbol{\theta})\right] + \mathbb{E}_{q(\boldsymbol{\theta})} \left[\log p(\boldsymbol{\theta})\right] - \mathbb{E}_{q(\boldsymbol{\theta})} \left[\log q(\boldsymbol{\theta})\right] + \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta}|\boldsymbol{y})}{p(\boldsymbol{\theta})} \; d\boldsymbol{\theta} \\
&= \mathbb{E}_{q(\boldsymbol{\theta})} \left[\log p(\boldsymbol{y}|\boldsymbol{\theta})\right] + \mathbb{E}_{q(\boldsymbol{\theta})} \left[\log p(\boldsymbol{\theta})\right] - \mathbb{E}_{q(\boldsymbol{\theta})} \left[\log q(\boldsymbol{\theta})\right] + D(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta}|\boldsymbol{y})) \qquad (3.12) \\
&= \mathbb{E}_{q(\boldsymbol{\theta})} \left[\log p(\boldsymbol{y}|\boldsymbol{\theta})\right] - D(q(\boldsymbol{\theta}) \| p(\boldsymbol{\theta})) + D(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta}|\boldsymbol{y})). \qquad (3.13)
\end{aligned}$$

Now as $D(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta}|\boldsymbol{y}))$ is positive, dropping the term from (3.13) gives a lower bound on the log model evidence.

$$\log p(\boldsymbol{y}) \geq \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log p(\boldsymbol{y}|\boldsymbol{\theta})\right] - D(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta})). \tag{3.14}$$

Now substituting in (3.3)

$$\mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{\theta})] - D(p(\boldsymbol{\theta}|\boldsymbol{y}) \parallel p(\boldsymbol{\theta})) \geq \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log p(\boldsymbol{y}|\boldsymbol{\theta})\right] D(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta})).$$

This then gives an upper bound on the penalty term:

$$D(p(\boldsymbol{\theta}|\boldsymbol{y}) \parallel p(\boldsymbol{\theta})) \leq \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{\theta})] - \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log p(\boldsymbol{y}|\boldsymbol{\theta})\right] + D(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta})). \tag{3.15}$$

The upper bound on the penalty term will be tight when $D(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta}|\boldsymbol{y}))$ is small. This is the same gap that appears in the usual variational evidence lower bound.

3.2.5 Sandwiching the evidence

Substituting the lower bound on the penalty term (3.10) into (3.3) gives an upper bound on the evidence:

$$\log p(\boldsymbol{y}) \leq \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{y})}\left[\log p(\boldsymbol{y}|\boldsymbol{\theta})\right] + \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{y})}\left[\log p(\boldsymbol{\theta})\right] + \frac{d}{2} + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log(\det(\Sigma_{\boldsymbol{\theta}|\boldsymbol{y}})). \tag{3.16}$$

Substituting the upper bound on the penalty term (3.10) into (3.3) gives the usual variational lower bound on the model evidence

$$\log p(\boldsymbol{y}) \geq \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log p(\boldsymbol{y}|\boldsymbol{\theta})\right] + \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log p(\boldsymbol{\theta})\right] - \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log q(\boldsymbol{\theta})\right]. \tag{3.17}$$

If the posterior distribution is normally distributed, and we use a Gaussian variational distribution both the upper and lower bounds will be tight. As the integrated likelihood is invariant to reparameterisations it may be worth searching for effective normalising transforms of $\boldsymbol{\theta}$. We will argue that subsampling can be used to estimate the bounds efficiently. Before developing our methodology further we review important existing related work.

3.3 Related work

In this section we review on subsampling for tall datsets and existing methods for calculating the integrated likelihood. As mentioned in the introduction, the pseudo-marginal MCMC algorithm requires unbiased estimation of the likelihood from subsamples. Existing work builds an unbiased estimator, or asymptotically unbiased estimator of the likelihood from unbiased estimators of the log likelihood (Quiroz et al., 2016; Bardenet et al., 2017; Quiroz et al., 2018). We focus on Monte Carlo estimators of the integrated likelihood that can be implemented given posterior samples and some generic closed form approximations. More sophisticated estimators of the integrated likelihood such as nested sampling (Skilling et al., 2006) or the power posterior method (Friel and Pettitt, 2008) are not considered here as they require more specialised MCMC implementations.

3.3.1 Subsampled log likelihoods

Unbiased estimation of the likelihood using subsamples is difficult as the likelihood is a product of n terms. Unbiased estimation of the log likelihood, $\ell(\boldsymbol{\theta}) = \log p(\boldsymbol{y}|\boldsymbol{\theta})$, is significantly easier as $\log p(\boldsymbol{y}|\boldsymbol{\theta})$ is a sum over n log likelihood contributions,

$$\log p(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\boldsymbol{y}_i|\boldsymbol{\theta}).$$

The simplest approach is to take a uniform random sample of size m, where we assume $m \ll n$. Let S denote the set of subsampled indices, so $S \subset \{1, \ldots, n\}$ and $|S| = m$. Then a simple estimator of the log likelihood at $\boldsymbol{\theta}$ is

$$\widehat{\ell}_{\text{simple}}(\boldsymbol{\theta}) = \frac{n}{m} \sum_{i \in S} \log p(y_i|\boldsymbol{\theta}).$$

Although easy to understand and implement, this estimator scales very poorly with n. Let $v_n(\boldsymbol{\theta})$ denote the population variance of the log likelihood values evaluated at $\boldsymbol{\theta}$ over the n observations. The variance of the estimator is seen to be

$$\text{var } \widehat{\ell}_{\text{simple}}(\boldsymbol{\theta}) = \frac{n^2}{m} v_n(\boldsymbol{\theta}).$$

It is reasonable to expect $v_n(\boldsymbol{\theta})$ to stabilise around a constant as n increases. Thus, in order to control the variance as n increases, we have to increase the batch size m quadratically with n. This is undesirable as we would like the computational cost of estimating the integrated likelihood to be sublinear in n. To reduce the variance of the log likelihood estimator we can employ control variates, a generic technique for reducing Monte Carlo variance. Control variates have been identified as a critical tool in many Big Data subsampling algorithms (Baker et al., 2017). Suppose we have a Monte Carlo estimator Z such that $\mathbb{E}[Z] = \mu$, where μ is the unknown quantity of interest. Suppose that we can define another random variable W on the same probability space with known expectation τ. Let α denote some constant. We can define a new estimator Z_{CV} that makes use of the auxiliary control variate W:

$$Z_{CV} = Z + \alpha(W - \tau).$$

The estimator Z_{CV} is also an unbiased estimator for μ for any choice of constant α. The variance of the new estimator is

$$\text{var}(Z_{CV}) = \text{var}(Z) + \alpha^2 \text{var}(W) + 2\alpha \times \text{cov}(Z, W). \tag{3.18}$$

If Z and W are correlated, and α is chosen appropriately, $\text{var}(Z_{CV})$ can be much smaller than $\text{var}(Z)$. The full dataset log likelihood at $\boldsymbol{\theta}$ is the sum of n contributions $\log p(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{i=1}^n \log p(\boldsymbol{y}_i|\boldsymbol{\theta}) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta})$. The difference estimator is a special control variate scheme for estimating population totals from a subsample (Sarndal et al., 1992). The difference estimator introduces an approximate log likelihood contribution for each observation $\widehat{\ell}_i(\boldsymbol{\theta})$ for $i = 1, \ldots, n$. As in Bardenet et al. (2017) and Quiroz et al. (2018) we consider the use of Taylor series approximations about the maximum likelihood estimate to form the log likelihood approximations $\widehat{\ell}_i(\boldsymbol{\theta})$ for $i = 1, \ldots, n$. The control variates can reduce the variance compared to the simple estimator $\widehat{\ell}_{\text{simple}}(\boldsymbol{\theta})$. Bardenet et al. and Quiroz et al. both consider second order approximations. We also consider a first order approximation, as computing the Hessian matrix can be challenging for high-dimensional statistical models.

Let $\widehat{\boldsymbol{\theta}}$ denote the maximum likelihood estimate of $\boldsymbol{\theta}$. Let $\boldsymbol{g}_i$ denote the gradient of the log likelihood evaluated at $\widehat{\boldsymbol{\theta}}$ for observation i. Likewise, let $\boldsymbol{H}_i$ denote the Hessian matrix of the log likelihood contribution of observation i, evaluated at the maximum likelihood estimate. For $i = 1, \ldots, n$:

$$\boldsymbol{g}_i = \nabla \log p(\boldsymbol{y}_i|\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}, \quad \boldsymbol{H}_i = \nabla^2 \log p(\boldsymbol{y}_i|\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}.$$

Let $\boldsymbol{g}$ denote the full dataset gradient so $\boldsymbol{g} = \sum_{i=1}^n \boldsymbol{g}_i$. Let $\boldsymbol{H}$ give the full dataset Hessian matrix, so $\boldsymbol{H} = \sum_{i=1}^n \boldsymbol{H}_i$. For $i = 1, \ldots, n$ the first order approximation to the log likelihood contribution is given by

$$\widehat{\ell}_{i,1}(\boldsymbol{\theta}) = \ell(\widehat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^\intercal \boldsymbol{g}_i. \tag{3.19}$$

For $i = 1, \ldots, n$ the second order approximation to the log likelihood contribution is given by

$$\widehat{\ell}_{i,2}(\boldsymbol{\theta}) = \ell(\widehat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^\intercal \boldsymbol{g}_i + \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^\intercal \boldsymbol{H}_i (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}). \tag{3.20}$$

The first order log likelihood estimator $\widehat{\ell}_{\text{gradient}}(\boldsymbol{\theta})$ is defined as :

$$\widehat{\ell}_{\text{gradient}}(\boldsymbol{\theta}) = \log p(\boldsymbol{y}|\widehat{\boldsymbol{\theta}} + \frac{n}{m}\sum_{i\in S}\left[\ell_i(\boldsymbol{\theta}) - \widehat{\ell}_{i,1}(\boldsymbol{\theta})\right]. \tag{3.21}$$

The second order log likelihood estimator $\widehat{\ell}_{\text{hessian}}(\boldsymbol{\theta})$ is defined as:

$$\widehat{\ell}_{\text{hessian}}(\boldsymbol{\theta}) = \log p(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})^{\intercal}\boldsymbol{H}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}) + \frac{n}{m}\sum_{i\in S}\left[\ell_i(\boldsymbol{\theta}) - \widehat{\ell}_{i,2}(\boldsymbol{\theta})\right]. \tag{3.22}$$

The variance of $\widehat{\ell}_{\text{gradient}}$ and $\widehat{\ell}_{\text{hessian}}$ is a function of the remainder terms in the Taylor series approximations (3.19) and (3.20). Qualitatively, the smaller the remainder terms, the smaller the variance. We analyse the asymptotic variance of the estimators in the Appendix to this chapter. We make a heuristic argument under mild assumptions that the variance of $\widehat{\ell}_{\text{gradient}}$ is $O_p(m^{-1})$ and the variance of $\widehat{\ell}_{\text{hessian}}$ is $O_p(n^{-1})$. Stochastic notation is used to account for the fact that we take $\boldsymbol{\theta}$ to be a stochastic sequence, where we are treating $\boldsymbol{\theta}$ as a sample from a posterior distribution.

3.3.2 Subsampled likelihoods

For simplicity suppose we have n independent observations $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$. Unbiased estimation of the likelihood $p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^n p(\boldsymbol{y}_i|\boldsymbol{\theta})$ is difficult as it is a product over n terms. The Poisson estimator (Wagner, 1987; Papaspiliopoulos, 2009) builds an unbiased estimator of the likelihood from unbiased estimators of the log likelihood. We assume that we use one of the subsampling based esitmators of the log likelihood described in the previous subsection. Let $(\widehat{\ell}_j(\boldsymbol{\theta}))$ be a sequence of independently and identically distributed unbiased estimators of the log likelihood for $j \in \mathbb{N}$. The basic Poisson estimator is defined as the randomised product

$$\widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}) = \exp(\lambda)\prod_{j=1}^{J}\frac{\widehat{\ell}_j(\boldsymbol{\theta})}{\lambda}, \quad \text{where } J \sim \text{Poisson}(\lambda). \tag{3.23}$$

Taking iterated expectations shows the estimator is unbiased. We take expectations over the subsampling indices for the log likelihood estimators S, and the random number of product terms J.

$$\begin{aligned}
\mathbb{E}_{S,J}[\widehat{p}(\boldsymbol{y}|\boldsymbol{\theta})] &= \mathbb{E}_J[\mathbb{E}_{S|J}[\widehat{p}(\boldsymbol{y}|\boldsymbol{\theta})|J=j]] \\
&= \mathbb{E}_J\left[\frac{\exp(\lambda)}{\lambda^j}\prod_{k=1}^{j}\mathbb{E}_{S|J}[\widehat{\ell}_j(\boldsymbol{\theta})]\right] \\
&= \mathbb{E}_J\left[\frac{\exp(\lambda)}{\lambda^j}\prod_{k=1}^{j}\ell(\boldsymbol{\theta})\right] \\
&= \mathbb{E}_J\left[\frac{\exp(\lambda)}{\lambda^j}\ell(\boldsymbol{\theta})^j\right] \\
&= \sum_{j=0}^{\infty}\Pr(J=j)\frac{\exp(\lambda)}{\lambda^j}\ell(\boldsymbol{\theta})^j \\
&= \sum_{j=0}^{\infty}\frac{1}{j!}\exp(-\lambda)\lambda^j \times \frac{\exp(\lambda)}{\lambda^j}\ell(\boldsymbol{\theta})^j \\
&= \sum_{j=0}^{\infty}\frac{\ell(\boldsymbol{\theta})^j}{j!} \\
&= \exp(\ell(\boldsymbol{\theta})) \\
&= p(\boldsymbol{y}|\boldsymbol{\theta}).
\end{aligned}$$

The second line relies on the unbiasedness and the independence of the log likelihood estimators $\widehat{\ell}_1(\boldsymbol{\theta}), \ldots, \widehat{\ell}_j(\boldsymbol{\theta})$. The fourth line uses the probability mass function of the Poisson distribution. The seventh line uses the

power series expansion of $\exp(z) = \sum_{i=0}^{\infty} z^i/i!$. The value of λ sets the expected computational cost of the Poisson estimator. Assuming the unbiased log likelihood estimators use a subsample of size m, the expected number of likelihood evaluations per use is $m\lambda$.

In practice it is common to introduce an extra scalar tuning parameter ω to reduce the variance. The modified estimator is again a randomised product:

$$\exp(\lambda + \omega) \prod_{j=1}^{J} \frac{\widehat{\ell}_j(\boldsymbol{\theta}) - \omega}{\lambda}, \quad J \sim \text{Poisson}(\lambda). \tag{3.24}$$

The modified Poisson estimator is also unbiased. From the general result given by Papaspiliopoulos (2009), we have that for a fixed λ, the choice of ω giving the minimum variance is $\omega = \log p(\boldsymbol{y}|\boldsymbol{\theta}) - \lambda$. The minimum variance tuning parameter requires the unknown log likelihood at $\boldsymbol{\theta}$.

An important issue is that both the simple Poisson estimator and the modified Poisson estimator are possibly negative. The modified estimator is almost surely nonnegative if ω is chosen such that $\widehat{\ell}_j(\boldsymbol{\theta}) > \omega$ holds almost surely. It can be difficult to determine such a bound analytically, and a conservative choice of ω can inflate the Monte Carlo variance (Quiroz et al., 2018). A general result is that without additional information about the log likelihood there is no algorithm that takes unbiased estimators of the log likelihood and outputs an almost-surely nonnegative estimator of the likelihood (Jacob et al., 2015). An alternative subsampled likelihood estimator is the estimator proposed in Rhee and Glynn (2015). The Rhee and Glynn estimator also uses unbiased estimators of the log likelihood in a randomised product. A lower bound on the likelihood is also required to ensure a nonnegative estimator. In general it is difficult to construct unbiased and almost surely nonnonegative estimators of the likelihood that use subsampling.

3.3.3 Importance sampling

The model evidence can be viewed as the expected likelihood over the prior distribution $p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$. A direct approach for estimating the model evidence is to take W samples from the prior distribution $\boldsymbol{\theta}^{[1]}, \ldots, \boldsymbol{\theta}^{[W]}$, and to set

$$\widehat{p}(\boldsymbol{y}) = \frac{1}{W} \sum_{w=1}^{W} p(\boldsymbol{y}|\boldsymbol{\theta}^{[w]}). \tag{3.25}$$

Although simulation consistent, this strategy is typically highly inefficient (Raftery, 1996; Friel and Wyse, 2012). The likelihood will be concentrate in a small region supported by the prior, and as such a small fraction of the prior samples will dominate the sum. This leads to the estimator (3.25) having very high variance. Importance sampling estimators express the model evidence as an integral over some importance distribution $q(\boldsymbol{\theta})$ as opposed to an integral over the prior distribution $p(\boldsymbol{\theta})$. The identity at the heart of this strategy is

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \int \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Given R samples from $q(\boldsymbol{\theta})$, denoted $\boldsymbol{\theta}_q^{[1]}, \ldots, \boldsymbol{\theta}_q^{[R]}$, the importance sampling estimator is

$$\widehat{p}(\boldsymbol{y}) = \frac{1}{R} \sum_{r=1}^{R} \frac{p(\boldsymbol{y}|\boldsymbol{\theta}_q^{[r]})p(\boldsymbol{\theta}_q^{[r]})}{q(\boldsymbol{\theta}_q^{[r]})}. \tag{3.26}$$

If the importance distribution $q(\boldsymbol{\theta})$ is exactly equal to the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ then the Monte Carlo estimator (3.26) has zero variance. In this ideal scenario $q(\boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})/p(\boldsymbol{y})$ and the estimator

reduces to

$$\begin{aligned}\widehat{p}(\boldsymbol{y}) &= \frac{1}{R}\sum_{r=1}^{R}\frac{p(\boldsymbol{y}|\boldsymbol{\theta}^{[r]})p(\boldsymbol{\theta}_q^{[r]})}{q(\boldsymbol{\theta}_q^{[r]})} \\ &= \frac{1}{R}\sum_{r=1}^{R}\frac{p(\boldsymbol{y}|\boldsymbol{\theta}_q^{[r]})p(\boldsymbol{\theta}_q^{[r]})}{p(\boldsymbol{y}|\boldsymbol{\theta}_q^{[r]})p(\boldsymbol{\theta}_q^{[r]})}\frac{p(\boldsymbol{y})}{1} \\ &= \frac{Rp(\boldsymbol{y})}{R} \\ &= p(\boldsymbol{y}).\end{aligned}$$

Gelfand and Dey (1994) propose to set the importance distribution $q(\boldsymbol{\theta})$ to be a normal distribution with mean and covariance matrix given by the sample mean and covariance of the posterior samples $\boldsymbol{\theta}^{[1]}, \ldots, \boldsymbol{\theta}^{[B]}$. This is a popular choice in practice.

3.3.4 Harmonic mean estimator

The harmonic mean estimator (Newton and Raftery, 1994) uses posterior samples to estimate the marginal likelihood. The harmonic mean estimator is based on the identity

$$\begin{aligned}\mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{y})}\left[\frac{1}{p(\boldsymbol{y}|\boldsymbol{\theta})}\right] &= \int \frac{1}{p(\boldsymbol{y}|\boldsymbol{\theta})}p(\boldsymbol{\theta}|\boldsymbol{y})\ d\boldsymbol{\theta} \\ &= \int \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{y})}\ d\boldsymbol{\theta} \\ &= \frac{1}{p(\boldsymbol{y})}\int p(\boldsymbol{\theta})\ d\boldsymbol{\theta} \\ &= \frac{1}{p(\boldsymbol{y})}.\end{aligned}$$

Given B posterior samples $\boldsymbol{\theta}^{[1]}, \ldots, \boldsymbol{\theta}^{[B]}$, a simulation consistent estimator of the model evidence is then

$$\widehat{p}(\boldsymbol{y}) = \left(\frac{1}{B}\sum_{b=1}^{B} p(\boldsymbol{y}|\boldsymbol{\theta}^{[b]})\right)^{-1}.$$

An advantage of the harmonic mean estimator is that it is readily computable if the posterior samples were generated using a Metropolis-Hastings algorithm (eg. Algorithm 3.1). If the likelihood ratio calculations are saved at each iteration, the harmonic mean estimator can be computed with very little additional cost after the posterior sampling run. In the tall data setting, it is perhaps unlikely that full likelihood evaluations have been made in the process of generating $\boldsymbol{\theta}^{[1]}, \ldots, \boldsymbol{\theta}^{[B]}$, so this benefit may not be present. The harmonic mean estimator can have infinite variance, and can exhibit a large finite sample bias (Friel and Wyse, 2012).

3.3.5 Bridge sampling

Bridge sampling (Bennett, 1976; Meng and Wong, 1996) also uses an importance distribution $q(\boldsymbol{\theta})$ but is based on a slightly more complicated identity. We first start with the relationship

$$1 = \frac{\int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})a(\boldsymbol{\theta})q(\boldsymbol{\theta})\ d\boldsymbol{\theta}}{\int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})a(\boldsymbol{\theta})q(\boldsymbol{\theta})\ d\boldsymbol{\theta}},$$

for some bridge function $a(\boldsymbol{\theta})$. Multiplying both sides of the equation by $p(\boldsymbol{y})$ gives

$$\begin{aligned}p(\boldsymbol{y}) &= \frac{\int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})a(\boldsymbol{\theta})q(\boldsymbol{\theta})\ d\boldsymbol{\theta}}{\int [p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})/p(\boldsymbol{y})]a(\boldsymbol{\theta})q(\boldsymbol{\theta})\ d\boldsymbol{\theta}} \\ &= \frac{\mathbb{E}_{q(\boldsymbol{\theta})}[p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})a(\boldsymbol{\theta})]}{\mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{y})}[a(\boldsymbol{\theta})q(\boldsymbol{\theta})]}.\end{aligned}$$

Suppose we have R samples from the importance density $q(\boldsymbol{\theta})$, denoted $\boldsymbol{\theta}_q^{[1]}, \ldots, \boldsymbol{\theta}_q^{[R]}$. Suppose we have B samples from the posterior distribution, $\boldsymbol{\theta}^{[1]}, \ldots, \boldsymbol{\theta}^{[B]}$

Bridge sampling requires the choice of importance distribution $q(\boldsymbol{\theta})$ and bridge function $a(\boldsymbol{\theta})$. There are some theoretical arguments to take $q(\boldsymbol{\theta})$ to resemble the posterior distribution. A common approach is to make a normal approximation. Let $s_1 = R/(R+B)$ and $s_2 = B/(R+B)$ The optimal choice of bridge function in terms of minimising mean square error is

$$a_{\text{opt}}(\boldsymbol{\theta}) = \frac{1}{s_1 p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) + s_2 p(\boldsymbol{y})q(\boldsymbol{\theta})}.$$

When setting $a(\boldsymbol{\theta}) = 1/q(\boldsymbol{\theta})$, bridge sampling reduces to ordinary importance sampling. The optimal bridge function is unattainable as it includes the unknown $p(\boldsymbol{y})$. Meng and Wong propose an iterative scheme to estimate the optimal bridge function, which in turn gives an iterative estimator of $p(\boldsymbol{y})$. The iterative bridge sampling estimator is

$$\widehat{p}(\boldsymbol{y})^{(t+1)} = \frac{R^{-1}\sum_{r=1}^{R} \dfrac{p(\boldsymbol{y}|\boldsymbol{\theta}_q^{[r]})p(\boldsymbol{\theta}_q^{[r]})}{s_1 p(\boldsymbol{y}|\boldsymbol{\theta}_q^{[r]})p(\boldsymbol{\theta}_q^{[r]}) + s_2 \widehat{p}(\boldsymbol{y})^{(t)} q(\boldsymbol{\theta}_q^{[r]})}}{B^{-1}\sum_{b=1}^{B} \dfrac{q(\boldsymbol{\theta}^{[b]})}{s_1 p(\boldsymbol{y}|\boldsymbol{\theta}^{[b]})p(\boldsymbol{\theta}^{[b]}) + s_2 \widehat{p}(\boldsymbol{y})^{(t)} q(\boldsymbol{\theta}^{[b]})}}. \tag{3.27}$$

The iterative scheme requires full likelihood evaluations for the B posterior samples, and the R samples from the importance distribution. The iterative estimator (3.27) has an interpretation as a maximum profile likelihood estimator of the model evidence under the assumption of independent posterior samples (Geyer, 1996; Shirts et al., 2003).

3.3.6 Laplace approximation

The Laplace approximation can be motivated under the assumption that the posterior density is approximately quadratic around the posterior mode $\widetilde{\boldsymbol{\theta}}$. Let $\boldsymbol{G}$ and $\boldsymbol{H}$ give the gradient and Hessian matrix of the unnormalised posterior evaluated at the mode respectively:

$$\begin{aligned}\boldsymbol{G} &= \nabla[\log p(\boldsymbol{y}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})]\big|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}} \\ \boldsymbol{H} &= \nabla^2[\log p(\boldsymbol{y}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})]\big|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}}.\end{aligned}$$

A second order Taylor expansion around the mode gives

$$\begin{aligned}p(\boldsymbol{y}) &\approx \int \exp\left[\log p(\boldsymbol{y}|\widetilde{\boldsymbol{\theta}}) + \log p(\widetilde{\boldsymbol{\theta}}) + (\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})^\intercal \boldsymbol{G} + \frac{1}{2}(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})^\intercal \boldsymbol{H}(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})\right] d\boldsymbol{\theta} \\ &= \int \exp\left[\log p(\boldsymbol{y}|\widetilde{\boldsymbol{\theta}}) + \log p(\widetilde{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})^\intercal \boldsymbol{H}(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})\right] d\boldsymbol{\theta} \\ &= \exp\left[\log p(\boldsymbol{y}|\widetilde{\boldsymbol{\theta}}) + \log p(\widetilde{\boldsymbol{\theta}})\right] \int \exp\left[\frac{1}{2}(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})^\intercal \boldsymbol{H}(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})\right] d\boldsymbol{\theta}.\end{aligned}$$

The second line uses the fact that gradient is zero at the posterior mode. Let $\boldsymbol{S}^{-1} = (-\boldsymbol{H})$ so that the integral can be recognised as the kernel of a multivariate Gaussian density with covariance matrix $\boldsymbol{S}$. Introducing the required normalising constant and integrating over the density function yields the approximation

$$\begin{aligned}p(\boldsymbol{y}) &\approx \exp\left[\log p(\boldsymbol{y}|\widetilde{\boldsymbol{\theta}}) + \log p(\widetilde{\boldsymbol{\theta}})\right] \int \exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})^\intercal \boldsymbol{S}^{-1}(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})\right] d\boldsymbol{\theta} \\ &= \exp\left[\log p(\boldsymbol{y}|\widetilde{\boldsymbol{\theta}}) + \log p(\widetilde{\boldsymbol{\theta}})\right] (2\pi)^{d/2}|\boldsymbol{S}|^{1/2} \int \frac{1}{(2\pi)^{d/2}|\boldsymbol{S}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})^\intercal \boldsymbol{S}^{-1}(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})\right] d\boldsymbol{\theta} \\ &= \exp\left[\log p(\boldsymbol{y}|\widetilde{\boldsymbol{\theta}}) + \log p(\widetilde{\boldsymbol{\theta}})\right] (2\pi)^{d/2}|\boldsymbol{S}|^{1/2}.\end{aligned}$$

The Laplace approximation to the log model evidence is

$$\log p(\boldsymbol{y}) \approx \log p(\boldsymbol{y}|\widetilde{\boldsymbol{\theta}}) + \log p(\widetilde{\boldsymbol{\theta}}) + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log(\det(\boldsymbol{S})). \tag{3.28}$$

Although the Laplace approximation has been shown to be competitive in simulation studies, it can be difficult to bound the error in the approximation in practice (Gelfand and Dey, 1994). Additionally, computation of the Hessian matrix can be difficult in high-dimensional statistical models.

3.3.7 Laplace-Metropolis estimator

The Laplace-Metropolis estimator (Raftery, 1996; Lewis and Raftery, 1997) is a variant of the Laplace approximation designed to be used more easily in situations where posterior samples are available. The Laplace approximation requires the posterior mode and the Hessian evaluated at the mode. These are not immediately available given posterior samples $\boldsymbol{\theta}^{[1]}, \ldots, \boldsymbol{\theta}^{[B]}$. Raftery proposes a number of candidate estimators for the posterior mode $\widetilde{\boldsymbol{\theta}}$ using the posterior samples:

1. Take $\widetilde{\boldsymbol{\theta}}$ to be the posterior sample that maximises $p(\boldsymbol{y}|\boldsymbol{\theta}^{[b]})p(\boldsymbol{\theta}^{[b]})$ over $b = 1, \ldots, B$.
2. Estimate the components of $\widetilde{\boldsymbol{\theta}}$ by taking the 1 component wise means of the posterior samples.
3. Estimate the components of $\widetilde{\boldsymbol{\theta}}$ by taking the component wise medians of the posterior samples.
4. Estimate $\widetilde{\boldsymbol{\theta}}$ by finding the multivariate median over the samples.

Let $\Sigma_{\boldsymbol{\theta}|\boldsymbol{y}}$ denote the covariance matrix of the posterior distribution. Raftery suggests using $\Sigma_{\boldsymbol{\theta}|\boldsymbol{y}}$ in place of $\boldsymbol{S} = (-\boldsymbol{H})^{-1}$ in the Laplace approximation (3.28). This is justified using an asymptotic argument. The posterior covariance matrix $\Sigma_{\boldsymbol{\theta}|\boldsymbol{y}}$ can be estimated using the empirical covariance matrix of the posterior samples or an alternative robust covariance matrix estimator. The Laplace-Metropolis estimator is

$$\log p(\boldsymbol{y}) \approx \log p(\boldsymbol{y}|\boldsymbol{\theta}^*) + \log p(\boldsymbol{\theta}^*) + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log(\det(\widehat{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{y}})), \tag{3.29}$$

where $\boldsymbol{\theta}^*$ and $\widehat{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{y}}$ are suitable estimators of the posterior mode and posterior covariance matrix, which can be obtained using the posterior samples. The Laplace-Metropolis estimator has performed well in simulations but its theoretical properties are not well understood (Raftery, 1996).

3.4 Application to tall datasets

Importance sampling, the harmonic mean estimator and bridge sampling all require repeated full likelihood evaluations $p(\boldsymbol{y}|\boldsymbol{\theta})$. If each likelihood evaluation is $O(n)$, these techniques may be impractical for large n datasets. As discussed in Section 3.3.2 subsampling can be used to provide unbiased estimates of likelihoods $\widehat{p}(\boldsymbol{y}|\boldsymbol{\theta})$. In principle it is possible to modify existing estimators of the model evidence to use subsampled likelihoods. Consider the basic importance sampling estimator described in Section 3.3.3. Given R samples from the importance distribution $q(\boldsymbol{\theta})$, denoted $\boldsymbol{\theta}_q^{[1]}, \ldots, \boldsymbol{\theta}_q^{[R]}$ we could replace the full likelihood evaluations $p(\boldsymbol{y}|\boldsymbol{\theta}^{[r]})$ with unbiased subsampled estimates $\widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}^{[r]})$ using the modified Poisson estimator in Section 3.3.2. The subsampled importance sampling estimator could be constructed as

$$\widehat{p}(\boldsymbol{y}) = \frac{1}{R}\sum_{r=1}^{R}\frac{\widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}^{[r]})p(\boldsymbol{\theta}^{[r]})}{q(\boldsymbol{\theta}^{[r]})}. \tag{3.30}$$

An issue is that we could obtain negative estimates of the model evidence unless the tuning parameter ω is an appropriate lower bound. Assuming that we plug in subsampled estimates of the likelihood, the possibility of negative estimates will also affect subsampled implementations of the harmonic mean estimator and subsampled bridge sampling. Although the Poisson estimator can cut the number of likelihood evaluations required to use the importance sampling estimators, the difficulties in choosing an appropriate lower bound ω make such an approach unattractive.

The Laplace approximation and the Laplace-Metropolis estimator are attractive alternatives that avoid much of the computational expense associated with the other methods. A drawback is that the

error associated with these approximate methods can be difficult to quantify in practice. As mentioned earlier, the Laplace approximation requires the posterior mode and Hessian, which is not automatically provided given posterior samples. The Laplace-Metropolis estimator approximates the mode and Hessian using information that is more readily available given simulation output. Although convenient, it is again difficult to determine the error that this introduces into the final estimator.

To integrate subsampling with the pseudo-marginal MCMC algorithm it is a requirement to have an unbiased estimator of the likelihood, necessitating the use of the Poisson estimator or alternatives. As our goal is Bayesian model selection rather than posterior sampling, we have the freedom to focus on the log model evidence. As already discussed, estimating $\log p(\boldsymbol{y}|\boldsymbol{\theta})$ using subsampling is significantly easier than estimating $p(\boldsymbol{y}|\boldsymbol{\theta})$ using subsampling.

At this point the computational advantages of the evidence bounds become more evident. As they are defined on the log scale it relatively easily to incorporate subsampling methodology. Suppose we have B samples from the posterior distribution. Let $\widehat{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{y}}$ be the empirical covariance matrix of the posterior samples $\boldsymbol{\theta}^{[1]}, \ldots, \boldsymbol{\theta}^{[B]}$. A simulation consistent estimate of the entropy upper bound is

$$\frac{1}{B}\sum_{i=1}^{B} \log p(\boldsymbol{y}|\boldsymbol{\theta}^{[i]}) + \frac{1}{B}\sum_{i=1}^{B} \log p(\boldsymbol{\theta}^{[b]}) + \frac{d}{2} + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log(\det(\widehat{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{y}})) \tag{3.31}$$

Suppose we generate R samples from the approximate variational posterior $q(\boldsymbol{\theta})$. Similar to the Gelfand and Dey (1994) recommendation for importance sampling, we propose to set the variational distribution $q(\boldsymbol{\theta})$ to be a normal distribution $N(\widehat{\boldsymbol{\mu}}_{\boldsymbol{\theta}|\boldsymbol{y}}, \widehat{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{y}})$ where $\widehat{\boldsymbol{\mu}}_{\boldsymbol{\theta}|y}$ and $\widehat{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{y}}$ are the mean and covariance matrix of the posterior samples $\boldsymbol{\theta}^{[1]}, \ldots, \boldsymbol{\theta}^{[B]}$. Let these samples be denotes $\boldsymbol{\theta}_q^{[1]}, \ldots, \boldsymbol{\theta}_q^{[R]}$. A simulation consistent estimator of the evidence lower bound is

$$\frac{1}{R}\sum_{r=1}^{R} \log p(\boldsymbol{y}|\boldsymbol{\theta}_q^{[r]}) + \frac{1}{R}\sum_{r=1}^{R} p(\boldsymbol{\theta}_q^{[q]}) + \frac{1}{R}\sum_{r=1}^{R} \log q(\boldsymbol{\theta}^{[q]}). \tag{3.32}$$

In some situations the expectations $\mathbb{E}_{q(\boldsymbol{\theta})}[\log p(\boldsymbol{\theta})]$ and $\mathbb{E}_{q(\boldsymbol{\theta})}[\log q(\boldsymbol{\theta})]$ can be determined analytically. This is the case when both the prior $p(\boldsymbol{\theta})$ and the variational approximation $\boldsymbol{q}(\boldsymbol{\theta})$ are normal distributions. The simulation consistent estimator of the variational lower bound can be simplified to

$$\frac{1}{R}\sum_{r=1}^{R} \log p(\boldsymbol{y}|\boldsymbol{\theta}_q^{[r]}) + \mathbb{E}_{q(\boldsymbol{\theta})}[\log p(\boldsymbol{\theta})] - \mathbb{E}_{q(\boldsymbol{\theta})}[\log q(\boldsymbol{\theta})]. \tag{3.33}$$

For tall datasets the B full log likelihood evaluations in (3.31) and the R full log likelihood evaluations in (3.32) and (3.33) will be computationally demanding. Subsampling can be used to estimate the log likelihood terms efficiently.

3.4.1 Estimation of evidence bounds

Suppose we have B samples from the posterior distribution $\boldsymbol{\theta}^{[1]}, \ldots, \boldsymbol{\theta}^{[B]}$ using one of the Big Data posterior simulation algorithms mentioned in the introduction. Pseudo-marginal MCMC or a stochastic gradient Langevin dynamics based sampler are two possibilities. To estimate the bounds (3.16) we need to estimate the posterior expectation $\mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{\theta})]$. We propose to use the subsampling estimators to estimate the log likelihoods. That is for some batch size $m \ll n$, to use either of the estimators

$$\widehat{\mathbb{E}}_{p(\boldsymbol{\theta}|\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{\theta})] = \frac{1}{B}\sum_{b=1}^{B} \widehat{\ell}_{\text{gradient}}(\boldsymbol{\theta}^{[b]}), \tag{3.34}$$

$$\widehat{\mathbb{E}}_{p(\boldsymbol{\theta}|\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{\theta})] = \frac{1}{B}\sum_{b=1}^{B} \widehat{\ell}_{\text{hessian}}(\boldsymbol{\theta}^{[b]}). \tag{3.35}$$

Similarly, we can use the control variates to estimate the required goodness of fit term $\mathbb{E}_{q(\boldsymbol{\theta})}[\log p(\boldsymbol{y}|\boldsymbol{\theta})]$ in the variational lower bound (3.14). Suppose R samples from the variational distribution $q(\boldsymbol{\theta})$ are

available, denoted $\boldsymbol{\theta}_q^{[1]}, \ldots, \boldsymbol{\theta}_q^{[R]}$. We propose to use either of the estimators

$$\widehat{\mathbb{E}}_{q(\boldsymbol{\theta})}[\log p(\boldsymbol{y}|\boldsymbol{\theta})] = \frac{1}{R}\sum_{r=1}^{R} \widehat{\ell}_{\text{gradient}}(\boldsymbol{\theta}_q^{[r]}), \tag{3.36}$$

$$\widehat{\mathbb{E}}_{q(\boldsymbol{\theta})}[\log p(\boldsymbol{y}|\boldsymbol{\theta})] = \frac{1}{R}\sum_{r=1}^{R} \widehat{\ell}_{\text{hessian}}(\boldsymbol{\theta}_q^{[r]}). \tag{3.37}$$

Given the subsampled estimators of the log likelihood we can define the subsampled estimator of the upper bound

$$\widehat{\log}\ p(\boldsymbol{y}) \leq \widehat{\mathbb{E}}_{p(\boldsymbol{\theta}|\boldsymbol{y})}\left[\log p(\boldsymbol{y}|\boldsymbol{\theta})\right] + \widehat{\mathbb{E}}_{p(\boldsymbol{\theta}|\boldsymbol{y})}\left[\log p(\boldsymbol{\theta})\right] + \frac{d}{2} + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log(\det(\widehat{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{y}})). \tag{3.38}$$

We can also define a subsampled estimator of the lower bound

$$\widehat{\log}\ p(\boldsymbol{y}) \geq \widehat{\mathbb{E}}_{q(\boldsymbol{\theta})}\left[\log p(\boldsymbol{y}|\boldsymbol{\theta})\right] + \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log p(\boldsymbol{\theta})\right] - \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log q(\boldsymbol{\theta})\right]. \tag{3.39}$$

3.5 Data application: flights dataset

We analyse the flights dataset that was also considered in Chapter 2. The flights dataset is available in the R package `nycflights13` (Wickham, 2014). There are $n = 327,346$ observations on flights departing New York City in 2013. There are data on 16 different carriers (airlines). We dichotomised the arrival delay variable (original units in minutes) to obtain a binary outcome. We labelled flights as late if the arrival delay was greater than zero, and on time if the arrival delay was less than or equal to zero. We compared three different logistic regression models for late arrival.

$$\mathcal{M}_1 = \text{intercept} : \text{carrier} + \text{delay} : \text{carrier}. \tag{3.40}$$

$$\mathcal{M}_2 = \text{intercept} : \text{carrier} + \text{delay} : \text{carrier} + \text{weekday} \tag{3.41}$$

$$\mathcal{M}_3 = \text{intercept} : \text{carrier} + \text{delay} : \text{carrier} + \text{weekday} : \text{carrier} \tag{3.42}$$

Model 1 is the unpooled model from Chapter 2. Model 2 introduces a fixed effect for the day of the week. It could be the case that flights are more likely to arrive late on a weekend. Model 3 allows for an interaction between weekday and carrier. We used independent $N(0,1)$ priors on all coefficients. We used a Gibbs sampler to generate $B = 5000$ observations from the full dataset posterior distribution. This dataset is of moderate size, and the covariates can be grouped for faster sampling. Ideally we would use a Big Data algorithm to generate the posterior samples, but we were unable to find an R package for doing so. We have not been concerned with the method of generation of the posterior samples, as our interest has been the follow up likelihood cost for then obtaining the model evidence. Our conclusions should not be affected by the fact that we have used a conventional Gibbs sampler to generate the initial collection of posterior samples.

The goodness of fit of each model were summarised by computing an Receiver Operating Characteristic (ROC) curve and a reliability plot. We generated predicted probabilities of late arrival for each observation. We again take the posterior predictive mean as the predicted probability of late arrival. The posterior predictive mean for a response given covariates $\boldsymbol{x}_i$ is obtained by integrating over the posterior distribution of the coefficients

$$\mathbb{E}[y_i|\mathcal{M}] = \int p(y_i = 1|x_i, \boldsymbol{\beta}, \mathcal{M}) p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y}, \mathcal{M}). \tag{3.43}$$

The ROC curve plots the in-sample true positive rate against the false positive rate using the predicted probabilities. The ROC curve summarises the predictive ability of each classifier. The area under the curve (AUC) is a summary measure for the predictive performance. A perfect classifier has AUC of 1, random guessing gives an AUC of 0.5. ROC curves were generated using the R package `pROC` (Robin et al., 2011).

Good predictive performance does not mean a model correctly specified. Another useful diagnostic is a reliability plot. Suppose a group of observations has a predicted probability of late arrival of 0.2. We would expect 20 percent of the observed data points to be late arrivals in this data group. The reliability plot checks the goodness of fit of the model in terms of empirical event probabilities matching the predicted event probabilities. Observations are binned according to predicted probabilities, and the average predicted probability is compared to the empirical probability of late arrival within the bin. The reliability curves were generated using the R package `caret` (Kuhn, 2008).

Figure 3.1 shows the ROC curves for each model. The AUC for each model is given in the Figure legend. The models have very similar AUC scores. Model 2 outperforms model 1 by 0.0023 and model 3 outperforms model 2 by 0.0006. Model 2 and 3 have more free parameters than model 1 so it is not surprising that they have better classification performance. What is more interesting is how the differences in predictive ability translate into Bayes factors. Before looking at this in more detail it is worth studying the reliability plots shown in Figure 3.2. Reliability curves are shown for each of the three models. A perfectly calibrated probabilistic classifier would give a reliability curve falling along the identity line in a sufficiently large sample. All three models appear very well calibrated. By the diagnostics in Figures 3.1 and 3.2 it is difficult to make a clear ranking of the models. We now turn to the model evidence.

We first computed the model evidence for each model using the importance sampling method proposed by Gelfand and Dey (1994). The estimator is described in Section 3.3.3. We generated $R = 5000$ samples from the importance distribution. The estimator required $R = 5000$ full likelihood evaluations. Estimated log Bayes factors are reported in Table 3.2. Moving from left to right, both model 2 and model 3 are supported over model 1. There is evidence for a day of the week effect on the probability of late arrival. Model 2 is preferred over model 3 with $2 \log \mathcal{B}_{23} = 55$. The day of the week effect appears constant across carriers. Model 2 has fewer parameters, and gives very similar results to Model 3 as seen in Figures 3.1 and 3.2. The computed Bayes factors are much larger than what is considered in Table 3.1. Here we have $n = 327,346$. We expect the typical magnitude of log Bayes factors to increase as n grows under standard asymptotic theory. Large sample Bayes factors can be difficult to interpret in isolation. The Bayes factors provide extra information to discriminate between models over what is shown in Figure 3.1 and Figure 3.2, however it is difficult to give an intuitive measure of the strength of evidence.

We then estimated the model evidence using the subsampled bound estimators (3.39) and (3.38). We estimated the required goodness of fit expectations using the three log likelihood estimators $\widehat{\ell}_{\text{simple}}$, $\widehat{\ell}_{\text{gradient}}$ and $\widehat{\ell}_{\text{hessian}}$ discussed in Section 3.3.1. For the lower bound, we took the variational distribution $q(\boldsymbol{\theta})$ to be the same normal approximation that was used in the implementation of importance sampling. We set the subsample size at $m = 500$. Table 3.3 reports the estimated evidence bounds and the Monte Carlo standard errors. Standard errors were obtained using the default method in the `mcmcse` R package (Flegal et al., 2017). We also report the estimates of the log model evidence using the importance sampling (IS) method. Standard errors for importance sampling were obtained using the nonparametric bootstrap.

The results in Table 3.3 show the importance of using control variates when estimating the full log likelihood from a subsample. The standard errors for $\widehat{\ell}_{\text{simple}}$ are much higher than the other estimators for all three models. In particular the standard errors for the bounds on Models 2 and 3 are larger than the difference in log Bayes factors. We can not determine which model has a the highest evidence score using the simple estimator $\widehat{\ell}_{\text{simple}}$. We obtain more precise estimates of the bounds using the control variate estimators $\widehat{\ell}_{\text{gradient}}$ and $\widehat{\ell}_{\text{hessian}}$.

The estimated bounds are reasonably tight. The model evidence can generally be enclosed in an interval of approximately width one. This suggests the normal approximation to the posterior distribution is reasonable under each model. Interestingly, the standard error of $\widehat{\ell}_{\text{hessian}}$ is greater than the standard error of $\widehat{\ell}_{\text{gradient}}$ for model 3. This is contrary to what we expect from the asymptotic analysis of the estimators. The larger standard error could be a result of the sampling method. We have used simple

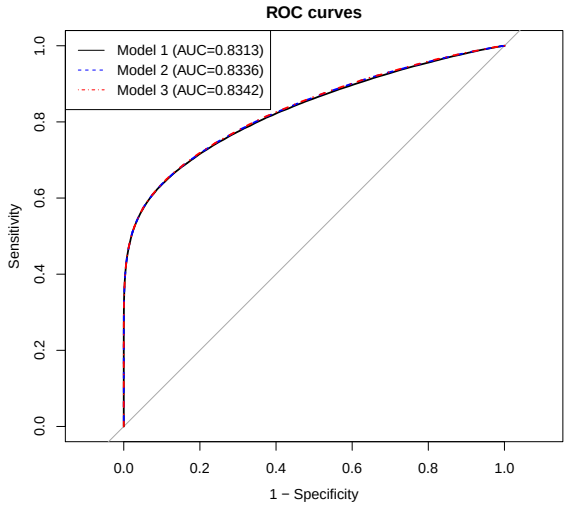


Figure 3.1: ROC curves for the flights dataset. Curves are shown for each of the three models in the candidate set. Classification performance is very similar across the three models. AUC values are reported in the legend. The addition of weekday improves in-sample predictive performance by a small margin.

random sampling in a situation where we do not have an equal number of observations for each carrier. The dummy coding in the design matrix will mean that a simple random subsample will contain very little information about the entire parameter set. We should aim to stratify the sample so that we get information about each parameter in the subsample. The second order estimator may be sensitive to this issue. The standard errors for the subsampled bound estimators are roughly an order of magnitude larger than the standard error of the importance sampling estimator. This should be taken into consideration when comparing the wall-clock time for each estimator. The standard errors in Table 3.3 are all estimated from a single replication of the computational experiment. Also we use different methodology to estimate the standard errors for the bounds and the importance sampling estimator. As future work we will repeat the experiment multiple times to obtain more robust estimates of the Monte Carlo standard errors.

Table 3.4 reports the time spent on likelihood evaluations for each estimator. We report the sum of the time to compute the upper and lower bounds for the subsampling based estimators. The subsampled estimators are multiple orders of magnitude faster than the standard importance sampling method. The subsampled estimators require $m = 500$ likelihood evaluations at each $\boldsymbol{\theta}$ with standard importance sampling requiring $n = 327,346$ likelihood evaluations. The second order estimator $\widehat{\ell}_{\text{hessian}}$ is considerably slower than the first order estimator $\widehat{\ell}_{\text{gradient}}$. The cost of the extra quadratic adjustment in (3.20) compared to (3.19) is non-negligible. As the second order estimator has higher standard errors, it seems the gradient based estimator is a better choice in this example.

We also implemented a subsampled version of importance sampling as per (3.30). We again generated $R = 5000$ samples from the same normal importance distribution that was used for the regular importance sampling method. We used the simple Poisson estimator (3.23) to generate unbiased estimates of the likelihood. The Poisson estimator requires unbiased estimators of the log likelihood as input. We tried the Poisson estimator with each of the log likelihood estimators $\widehat{\ell}_{\text{simple}}$, $\widehat{\ell}_{\text{gradient}}$ and $\widehat{\ell}_{\text{hessian}}$. We set $\lambda = 5$ and took $m = 100$ when using the subsampling estimators. This was so the expected number of likelihood evaluations per use was equal to five hundred. Table 3.5 reports the proportion of negative likelihood estimates obtained using each subsampling estimator. The proportion of negative estimates is close to 0.5 for each model and estimator. It is hard to construct a good estimator of the log model evidence given many negative likelihood estimates. It is difficult to use to simple Poisson estimator to accelerate importance samplers for the model evidence.

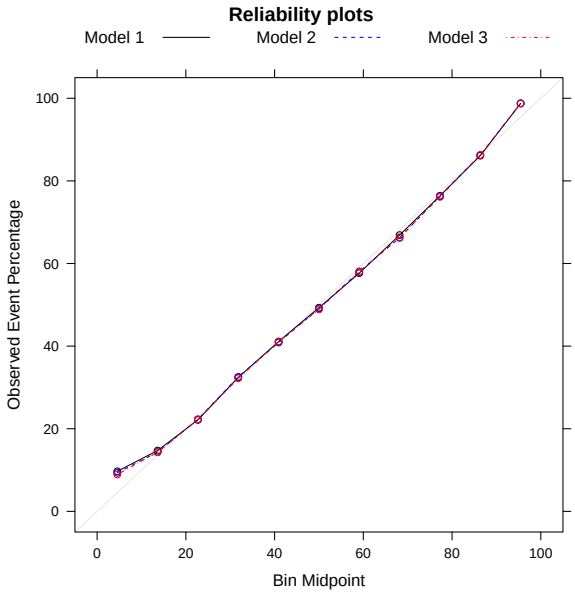


Figure 3.2: Reliability curves for the flights dataset. Curves are shown for each of the three models in the candidate set. All models are very well calibrated across the range of theoretical probabilities. Each model could be considered to be correctly specified.

	$2\log\mathcal{B}_{21}$	$2\log\mathcal{B}_{31}$	$2\log\mathcal{B}_{23}$
Value	941	887	55

Table 3.2: Log Bayes Factors for the flights dataset ($n = 327,346$). The values of the Bayes factors are all outside the range considered in Table 3.1. The models in the candidate set appear give very similar predictions and are all well calibrated. Interpreting the strength of evidence when n is large can be difficult.

	$\widehat{\ell}_{\text{simple}}$		$\widehat{\ell}_{\text{gradient}}$		$\widehat{\ell}_{\text{hessian}}$		IS
	Lower	Upper	Lower	Upper	Lower	Upper	
Model 1	-147209.1 (98.02)	-147100.8 (99.46)	-147112.5 (0.21)	-147111.9 (0.16)	-147112.6 (0.25)	-147111.9 (0.63)	-147112.0 (0.02)
Model 2	-146548.8 (122.18)	-146637.6 (97.23)	-146641.2 (0.22)	-146641.2 (0.19)	-146640.7 (0.76)	-146641.8 (0.20)	-146641.0 (0.02)
Model 3	-146491.5 (106.59)	-146611.4 (93.61)	-146669.6 (0.55)	-146668.5 (0.53)	-146579.6 (21.04)	-146588.0 (17.07)	-146668.7 (0.04)

Table 3.3: Estimates of the model evidence for the flights dataset. The first three columns report the evidence bounds obtained using the different log likelihood estimators. Estimated standard errors are in brackets. The Monte Carlo variance of the simple log likelihood estimator is too large to rank models confidently. The control variate based estimators have significantly lower standard errors. The widths of the evidence bounds are small in relation to the estimated Bayes factors between models.

	$\widehat{\ell}_{\text{simple}}$	$\widehat{\ell}_{\text{gradient}}$	$\widehat{\ell}_{\text{hessian}}$	IS
Model 1	3	4	10	523
Model 2	4	4	12	603
Model 3	12	13	85	1743

Table 3.4: Time spent on likelihood evaluations for the flights dataset (seconds). The importance sampling method requires a full $O(n)$ likelihood evaluation for each posterior sample. The subsampling based estimators require $O(m)$ likelihood evaluations per sample. We report the total time for calculating the lower and upper bounds for the subsampling estimators. In this simulation $n = n = 327,346$ and $m = 500$. We report the time spent on likelihood evaluations only. The time spent on generating subsampling integer sets S was recorded separately.

	$\widehat{\ell}_{\text{simple}}$	$\widehat{\ell}_{\text{gradient}}$	$\widehat{\ell}_{\text{hessian}}$
Model 1	0.52	0.50	0.50
Model 2	0.49	0.50	0.51
Model 3	0.50	0.50	0.51

Table 3.5: Proportion of negative likelihood estimates using the simple Poisson estimator over the $B = 5000$ posterior samples. We applied the simple Poisson estimator using each of the subsampled log likelihood esitmators. The high fraction of negative likelihood estimates makes it difficult to construct a subsampled version of importance sampling for the purposes of estimating the model evidence. The evidence bounds are defined on the log scale not subject to a positivity requirement.

3.6 Conclusion

Much of the status quo in Bayesian computation is not well suited to Big Data. As mentioned in the introduction, procedures that require repeated $O(n)$ evaluations of the full dataset likelihood are unsustainable as datasets grow taller. Estimation of the integrated likelihood can be computationally demanding on tall datasets because of this likelihood burden. We proposed a method for estimating the integrated likelihood using subsampling that avoids repeated $O(n)$ likelihood evaluations. The relevance to model choice is that the proposed estimator of the model evidence can be calculated in less time than traditional importance sampling methods. Bayesian model choice can then proceed at a less glacial pace. Practitioners will inevitably perform a cost benefit analysis when deciding what procedure to use for model selection. If the computational cost of the Bayesian approach is very high, the scales will tip in favour of alternative approaches, whatever the benefits of a Bayesian analysis may be. The main objective of the thesis is to identify and investigate strategies that minimise the expense of the compute step in Box's loop (refer back to Figure 1.1 in Chapter 1). The point of this Chapter is to see how subsampling can be used to minimise the expense of the compute step when we want to critique multiple models on a large dataset using the fully Bayesian paradigm. To reiterate, we are trying to give an air of computational feasibility to Bayesian purism in the huge n setting.

Subsampling has been used in algorithms for generating posterior samples $p(\boldsymbol{y}|\boldsymbol{\theta})$ given a large dataset $\boldsymbol{y}$, however subsampling based estimation of the model evidence does not seem to have been explored in great depth. Subsampling methodology for posterior simulation makes use of subsampled estimates of the log likelihood. Integrating these estimators into existing importance sampling methods for calculating the evidence is difficult due to the possibility of negative estimates of the likelihood. We found that the log likelihood estimators can also be used for the purposes of efficient estimation of the model evidence in a different manner. Our starting point was an identity for the log model evidence:

$$\log p(\boldsymbol{y}) = \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{\theta})] - D(p(\boldsymbol{\theta}|\boldsymbol{y}) \parallel p(\boldsymbol{\theta})). \tag{3.44}$$

The goodness of fit term $\mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{\theta})]$ can be estimated efficiently using subsampling. Given posterior samples, we can bound the penalty term $D(p(\boldsymbol{\theta}|\boldsymbol{y}) \parallel p(\boldsymbol{\theta}))$ using a maximum entropy argument and a standard variational Bayes approach. Control variates are essential in order to control the Monte Carlo variance of the algorithm, this is a similar finding to other work on subsampling algorithms for Bayesian computation (Baker et al., 2017; Bierkens et al., 2016). We investigated both first order and second order Taylor series approximations as control variates. The first order estimator has attractive theoretical properties and performed competitively with the second order estimator in the simulations. Asymptotic properties of Bayes factors suggests that interval estimators of the log model evidence may be acceptable for ranking models in the large n regime.

The proposed methodology is closely related to the standard Laplace approximation, however there are some key differences. The Laplace approximation is based on a second order Taylor series expansion about the posterior mode. It can be difficult to determine the error in the Laplace approximation in practice

(Gelfand and Dey, 1994). We make a normal approximation to the posterior distribution. Assuming the parameter space is unconstrained, this provides an upper bound on the log model evidence using the maximum entropy property of the Gaussian distribution. Although our approach is more computationally demanding than the standard Laplace approximation or the Laplace-Metropolis estimator it does come with additional guarantees on the error in the approximation.

We have used the maximum entropy property of the normal distribution in $\mathbb{R}^d$, and assumed that the parameter space is unconstrained. Another way to achieve a closed form upper bound on the posterior entropy is to use general properties of the exponential family. We typically have closed form expressions for the entropy of a distribution in the exponential family. The maximum entropy distribution of a random variables X taking values in the interval $[0, \infty)$ subject to the constraint that $\mathbb{E}[X] = 1/\lambda$ is an exponential distribution with density $f(\boldsymbol{x}) = \lambda \exp(-\lambda x)$. Suppose we had a single parameter θ with support $[0, \infty)$ in our model. Given posterior samples we can form an estimate of $\mathbb{E}[\theta|\boldsymbol{y}]$. We could then form an upper bound on the posterior entropy using the maximum entropy property of the exponential distribution. Using exponential family distributions for the maximum entropy bound also suggests taking the same exponential family distributions for the variational approximation $q(\boldsymbol{\theta})$ for the lower bound. This is an interesting direction that helps to further distinguish our approach from the Laplace and Laplace-Metropolis estimators.

3.7 Appendix

3.7.1 Control variates

We have defined $\ell(\boldsymbol{\theta}) = \log p(\boldsymbol{y}|\boldsymbol{\theta})$ and $\ell_i(\boldsymbol{\theta}) = \log p(\boldsymbol{y}_i|\boldsymbol{\theta})$. The variance of the gradient estimator $\widehat{\ell}_{\text{gradient}}(\boldsymbol{\theta})$ (3.21) will be a function of the remainder terms from each of the individual log likelihood contributions. Let R^i_{gradient} denote the remainder term from the approximation for observation i, so

$$R^i_{\text{gradient}} = \ell_i(\boldsymbol{\theta}) - \widehat{\ell}_{i,1}(\boldsymbol{\theta}).$$

An alternative expression for (3.21) is then

$$\widehat{\ell}_{\text{gradient}}(\boldsymbol{\theta}) = \ell(\widehat{\boldsymbol{\theta}}) + \frac{n}{m}\sum_{i \in S} R^i_{\text{gradient}}.$$

Now as $\text{var}(Z) \leq \mathbb{E}(Z^2)$ for any random variable Z, the variance can be upper bounded in terms of the expected squared remainder term

$$\text{var}(\widehat{\ell}_{\text{gradient}}(\boldsymbol{\theta})) \leq \frac{n}{m}\sum_{i=1}^{n} |R^i_{\text{gradient}}|^2. \tag{3.45}$$

If the remainder terms are small the variance of the estimator will be small. The introduction of the first order control variates can decrease the Monte Carlo variance of the log likelihood estimator if the linear approximations to the log likelihood contributions are accurate. The variance of $\widehat{\ell}_{\text{hessian}}(\boldsymbol{\theta})$ estimator will also be a function of the remainder terms from each of the individual log likelihood approximations. Let R^i_{hessian} denote the remainder term from the approximation for observation i, so

$$R^i_{\text{hessian}} = \ell_i(\boldsymbol{\theta}) - \widehat{\ell}_{i,2}(\boldsymbol{\theta}).$$

The second order estimator can be expressed as

$$\widehat{\ell}_{\text{hessian}}(\boldsymbol{\theta}) = \log p(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}) + \frac{n}{m}\sum_{i \in S} R^i_{\text{hessian}}$$

Now as $\text{var}(Z) \leq \mathbb{E}(Z^2)$ for any random variable Z, the variance of $\widehat{\ell}_{\text{hessian}}(\boldsymbol{\theta})$ can be upper bounded in terms of the expected squared remainder term

$$\text{var}(\widehat{\ell}_{\text{hessian}}(\boldsymbol{\theta})) \leq \frac{n}{m}\sum_{i=1}^{n} |R^i_{\text{hessian}}|^2. \tag{3.46}$$

It is reasonable to expect the remainder terms from the second order approximation R^i_{hessian} to be smaller in magnitude than the remainders R^i_{gradient} from the first order approximation. As a consequence the variance of the second order estimator should be smaller than the variance of the first order estimator. We study the asymptotic variance of each estimator to compare them in a more precise manner.

3.7.2 Asymptotic variance

Here we consider the asymptotic variance of the proposed estimators $\widehat{\ell}_{\text{gradient}}(\boldsymbol{\theta})$ and $\widehat{\ell}_{\text{hessian}}(\boldsymbol{\theta})$ as the sample size n is taken to infinity. As mentioned earlier, the variance of the estimators will depend on the remainder terms from the Taylor approximations for each likelihood contribution. Recall that R^i_{gradient} denotes the remainder term for observation i from the first order Taylor expansion about the maximum likelihood estimate. Suppose that we can bound the second order derivatives of the log likelihood for each observation $i = 1, \ldots, n$. Given a d-dimensional parameter, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)^\mathsf{T}$, assume that for all $j, k \in \{1, \ldots, d\}$,

$$\left| \frac{\partial^2 \log p(\boldsymbol{y}_i|\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \right| \leq M_i,$$

for some constant M_i, and $i = 1, \ldots, n$. This is possible for a wide range of models in the exponential family. Then by the Taylor-Lagrange inequality we can form bounds on the remainder term

$$|R^i_{\text{gradient}}| \leq \frac{M_i}{2} \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1^2, \quad |R^i_{\text{gradient}}|^2 \leq \frac{M_i^2}{4} \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1^4.$$

Substituting the squared remainder bound into (3.45) gives an upper bound on the variance

$$\begin{aligned} \text{var}(\widehat{\ell}_{\text{gradient}}(\boldsymbol{\theta})) &\leq \frac{n}{m} \sum_{i=1}^{n} \frac{M_i^2}{4} \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1^4 \\ &= \frac{1}{m} \left(\sum_{i=1}^{n} \frac{1}{n} \frac{M_i^2}{4} \right) n^2 \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1^4. \end{aligned} \tag{3.47}$$

To determine the asymptotic variance of the estimators we take the evaluation point $\boldsymbol{\theta}$ to be a stochastic sequence. Roughly speaking we assume the posterior distribution concentrates in a ball of radius $\sqrt{n}$ around the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ as n increases. One way to motivate this is through the Bernstein-von Mises theorem (Van Der Vaart, 1998, Chapter 10). Under mild conditions we expect the posterior distribution to be asymptotically normal,

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \approx N(\widehat{\boldsymbol{\theta}}, n^{-1} \mathcal{I}_1(\widehat{\boldsymbol{\theta}})),$$

where $\mathcal{I}_1(\widehat{\boldsymbol{\theta}})$ is the Fisher information matrix for a single observation evaluated at the maximum likelihood estimate (Van Der Vaart, 1998, Chapter 10). The Fisher information for a single observation is defined as $\mathcal{I}(\boldsymbol{\theta}) = -\mathbb{E}[\nabla^2 \log p(\boldsymbol{y}_i|\boldsymbol{\theta})|\boldsymbol{\theta}]$, where the expectation is over the generative model with known parameter $\boldsymbol{\theta}$. We expect the posterior variance to be $O(n^{-1})$. Using properties of the folded normal distribution, each element of norm

$$\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1 = \sum_{i=1}^{d} |\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_i|$$

can be reasoned to be $O_p(n^{-1/2})$. As such we can take $\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1^4$ to be of the order $O_p(n^{-2})$. Plugging this deviation rate into the variance bound (3.47), we can cancel out the inflation by n that debilitates the simple likelihood estimator,

$$\begin{aligned} \text{var}(\widehat{\ell}_{\text{gradient}}(\boldsymbol{\theta})) &\leq \frac{1}{m} \left(\sum_{i=1}^{n} \frac{1}{n} \frac{M_i^2}{4} \right) n^2 \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1^4 \\ &\approx \frac{1}{m} \left(\sum_{i=1}^{n} \frac{1}{n} \frac{M_i^2}{4} \right) n^2 \times O_p(n^{-2}) \\ &\approx O_p(m^{-1}). \end{aligned}$$

Under mild assumptions the term inside the sum $\left(\sum_{i=1}^{n} n^{-1} M_i^2\right)$ can be assumed to approach constant as $n \to \infty$. As such we expect the variance of the gradient estimator to be $O_p(1/m)$ for large n. The approximation in the final line is to reflect that this is a heuristic argument. This is a significant improvement over the simple subsampling estimator where $\text{var}(\widehat{\ell}_{\text{simple}}(\boldsymbol{\theta}))$ is expected to be $O_p(n^2/m)$.

A similar analysis can be performed for the second order estimator $\widehat{\ell}_{\text{hessian}}(\boldsymbol{\theta})$. Let R^i_{hessian} denote the remainder term for observation i from the second order Taylor expansion about the maximum likelihood estimate. We assume that we can bound the third order partial derivatives of the log likelihood contribution for each observation. Specifically, suppose that for all $j, k, l \in \{1, \ldots, d\}$,

$$\left| \frac{\partial^3 \log p(\boldsymbol{y}_i|\boldsymbol{\theta})}{\partial\theta_j \partial\theta_k \theta_l} \right| \leq M_i,$$

for some constant M_i, and $i = 1, \ldots, n$. Then by the Taylor-Lagrange inequality we can form bounds on the remainder term (Walschap, 2015, Chapter 2),

$$|R^i_{\text{hessian}}| \leq \frac{M_i}{6} \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1^3, \quad |R^i_{\text{hessian}}|^2 \leq \frac{M_i^2}{36} \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1^6.$$

Substituting the squared remainder bound into (3.46) gives an upper bound on the variance

$$\begin{aligned} \text{var}(\widehat{\ell}_{\text{hessian}}(\boldsymbol{\theta})) &\leq \frac{n}{m} \sum_{i=1}^{n} \frac{M_i^2}{36} \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1^6 \\ &= \frac{1}{m} \left(\sum_{i=1}^{n} \frac{1}{n} \frac{M_i^2}{36} \right) n^2 \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1^6. \end{aligned}$$

We again use the fact that we expect $\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1$ to be $O_p(n^{-1/2})$. Making another heuristic argument for the asymptotic variance,

$$\begin{aligned} \text{var}(\widehat{\ell}_{\text{hessian}}(\boldsymbol{\theta})) &\leq \frac{1}{m} \left(\sum_{i=1}^{n} \frac{1}{n} \frac{M_i^2}{36} \right) n^2 \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1^6 \\ &\approx \frac{1}{m} \left(\sum_{i=1}^{n} \frac{1}{n} \frac{M_i^2}{36} \right) n^2 \times O_p(n^{-3}) \\ &\approx \frac{1}{m} \left(\sum_{i=1}^{n} \frac{1}{n} \frac{M_i^2}{36} \right) \times O_p(n^{-1}). \end{aligned}$$

Again assuming that $\sum_{i=1}^{n} n^{-1} M_i^2$ approaches a constant, we can reason that the variance of the second order estimator approaches zero as n tends to infinity.

It is interesting to compare the behaviour of $\widehat{\ell}_{\text{gradient}}(\boldsymbol{\theta})$ an $\widehat{\ell}_{\text{hessian}}(\boldsymbol{\theta})$. The assumption that $\boldsymbol{\theta}$ approaches $\widehat{\boldsymbol{\theta}}$ at rate $O_p(n^{-1/2})$ implies that the remainder terms vanish. This is counterbalanced by the scaling factor n/m that appears in both the definition of both estimators (equations (3.34)). For the $\widehat{\ell}_{\text{hessian}}(\boldsymbol{\theta})$, the remainder terms vanish at a sufficiently fast rate such that the variance tends to zero. For the $\widehat{\ell}_{\text{gradient}}(\boldsymbol{\theta})$ the remainder terms diminish at a rate that leads to stable finite variance.

Example: logistic regression

The logistic regression model satisfies the assumptions in the analysis of the asymptotic variance. Let y_i denote the binary response and $\boldsymbol{x}_i$ represent the column vector of covariates for observation i, for $i = 1, \ldots, n$. We assume that $y_i \sim \text{Bernoulli}(\sigma(\eta_i))$, where $\eta_i = \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}$ and $\sigma(z) = 1/(1 + \exp(-z))$. For the analysis of $\widehat{\ell}_{\text{gradient}}$ in Section 3.7.2 to apply, it is necessary to bound the second order partial derivatives. The Hessian matrix of each log likelihood contribution is

$$\nabla^2 \log p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\beta}) = -\widehat{y}_i(1 - \widehat{y}_i)\boldsymbol{x}_i \boldsymbol{x}_i^{\mathsf{T}},$$

where $\widehat{y}_i = \sigma(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta})$. As $0 \leq \widehat{y}_i \leq 1$, it holds that $\widehat{y}_i(1-\widehat{y}_i) < 1/4$. The absolute value of the second order partial derivatives then satisfies

$$\left|\frac{\partial^2 \log p(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\beta})}{\partial\beta_j\partial\beta_k}\right| \leq \frac{1}{4}\|\boldsymbol{x}_i\|_1^2,$$

for each observation $i = 1,\ldots,n$. The analysis of $\widehat{\ell}_{\text{hessian}}$ in Section 3.7.2 assumed bounded third order partial derivatives. The logistic regression model satisfies

$$\left|\frac{\partial^3 \log p(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\beta})}{\partial\beta_j\partial\beta_k\beta_l}\right| \leq \frac{1}{4}\|\boldsymbol{x}_i\|_1^3,$$

for each observation $i = 1,\ldots,n$.

4

Statistical properties of sketching algorithms

Summary

Sketching is a probabilistic data compression technique that has been largely developed in the computer science community. Numerical operations on big datasets can be intolerably slow; sketching algorithms address this issue by generating a smaller surrogate dataset. Typically, inference proceeds on the compressed dataset. Sketching algorithms generally use random projections to compress the original dataset and this stochastic generation process makes them amenable to statistical analysis. We argue that the sketched data can be modelled as a random sample, thus placing this family of data compression methods firmly within an inferential framework. In particular, we focus on the Gaussian, Hadamard and Clarkson-Woodruff sketches, and their use in single pass sketching algorithms for linear regression with huge n. We explore the statistical properties of sketched regression algorithms and derive new distributional results for a large class of sketched estimators. We develop confidence intervals for sketching estimators and bounds on the relative efficiency of different sketching algorithms. An important finding is that the best choice of sketching algorithm in terms of mean square error is related to the signal to noise ratio in the source dataset. Finally, we demonstrate the theory and the limits of its applicability on two real datasets.

4.1 Introduction

Sketching is a general probabilistic data compression technique designed for Big Data applications (Cormode, 2011). Even routine calculations can be prohibitively computationally expensive on massive datasets. Computation time can be reduced to an acceptable level by allowing for some approximation error in the results. Sketching algorithms relax the computational task by generating a compressed version of the original dataset which then serves as a surrogate for calculations. The compressed dataset is referred to as a sketch as it acts as a compact representation of the full dataset. Sketching algorithms use a randomised compression stage which makes them interesting from a statistical viewpoint. Sketching algorithms for linear regression have attracted significant attention in the numerical linear algebra and theoretical computer science communities (Woodruff, 2014; Mahoney, 2011). In this paper we investigate the statistical properties of sketched regression algorithms, a perspective which has received little attention up to now.

To describe sketched regression in more detail, we first assume the data consists of a n-length response vector $\boldsymbol{y}$ and a $n \times p$ matrix of covariates, $\boldsymbol{X}$ which is of full rank. It is assumed that $n > p$. The objective is to find the optimal least squares coefficients. Given sufficient computational resources, these could be computed exactly as

$$\boldsymbol{\beta}_F = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}.$$

The subscript F is used to indicate the connection to the full dataset. Only two quantities are needed in order to determine $\boldsymbol{\beta}_F$, the Gram matrix $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$, and the marginal associations $\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$. Calculation of

$\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ requires $O(np^2)$ operations while computation of $\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$ needs only $O(np)$ calculations. There are two broad methods for sketched regression, complete sketching and partial sketching. Complete sketching is based on approximating both $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ and $\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$, whereas partial sketching only approximates the Gram matrix.

Sketching algorithms use random linear mappings to reduce the size of the dataset from n to k observations. The random linear mapping can be represented as a $k \times n$ sketching matrix $\boldsymbol{S}$. Complete sketching generates a k-length sketched response vector $\widetilde{\boldsymbol{y}}$ and a $k \times p$ matrix of sketched predictors $\widetilde{\boldsymbol{X}}$. The sketched data are computed through the linear mappings $\widetilde{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$ and $\widetilde{\boldsymbol{X}} = \boldsymbol{S}\boldsymbol{X}$. Partial sketching only generates a $k \times p$ matrix of sketched covariates $\widetilde{\boldsymbol{X}}$. We again use the random mapping $\widetilde{\boldsymbol{X}} = \boldsymbol{S}\boldsymbol{X}$.

The complete sketching estimator, $\boldsymbol{\beta}_S$, is defined as the least squares coefficients using the sketched responses and predictors,

$$\boldsymbol{\beta}_S = (\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{y}}. \tag{4.1}$$

The partial sketching estimator, $\boldsymbol{\beta}_P$, is defined as

$$\boldsymbol{\beta}_P = (\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}. \tag{4.2}$$

The key difference between (4.1) and (4.2) is that the partial sketched estimator $\boldsymbol{\beta}_P$ is constructed using the exact marginal associations $\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$. Given the sketched data, computation of $\boldsymbol{\beta}_S$ or $\boldsymbol{\beta}_P$ requires only $O(kp^2)$ operations, compared with the $O(np^2)$ required for $\boldsymbol{\beta}_F$.

The estimand within a sketching algorithm is the optimal coefficient vector $\boldsymbol{\beta}_F$. Sketching algorithms have the property that given a fixed k, the approximation error $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_F\|_2$ or $\|\boldsymbol{\beta}_P - \boldsymbol{\beta}_F\|_2$ remains probabilistically bounded even as $n \to \infty$. Designing estimators for approximate computation with such properties is very difficult, and is a common goal in the development of techniques for Big Data (Bardenet and Maillard, 2015; Phillips, 2016). The favourable scaling properties of sketching algorithms are a critical factor in making them stand apart from simple subsampling approaches, where it can be difficult to establish universal worst case bounds for large n (Drineas et al., 2006; Ma et al., 2015). The fact that sketching algorithms provide finite k guarantees for arbitrarily large n is a major reason they have received so much attention in the computer science community.

There is a large literature concerned with designing appropriate distributions for the random sketching matrix $\boldsymbol{S}$. Our focus is on data-oblivious random projections, where the distribution of the sketching matrix is not a function of the source data $(\boldsymbol{y}, \boldsymbol{X})$. An example is the Gaussian sketch, where each element is independently distributed as a $N(0, 1/k)$ variate. We also consider the Hadamard sketch and the Clarkson-Woodruff sketch, random projections that exploit structure and sparsity for computational efficiency.

Most existing results on the accuracy of sketching are universal worst case bounds (Woodruff, 2014; Mahoney and Drineas, 2016). This is typical for randomised algorithms, however a more detailed error analysis can provide important insights (Halko et al., 2011). We investigate the statistical properties of $\boldsymbol{\beta}_P$ and $\boldsymbol{\beta}_S$ when using data oblivious sketches. An important finding is upper and lower bounds on the relative efficiency of complete sketching to partial sketching in terms of the signal to noise ratio in the source dataset. The statistical analysis also allows the construction of exact confidence intervals for the Gaussian sketch, and asymptotic confidence intervals for other random projections, paving the way for their wider use in the statistical community interested in Big Data methods.

We start by reviewing the existing literature on sketching algorithms before investigating the statistical properties in more detail. At its core, sketched regression is a randomised algorithm for approximate computation of $\boldsymbol{\beta}_F$. Repeated application of the sketching algorithm on the same dataset will produce different results. The first stage in our analysis is to establish the distributional properties of the sketched estimators with the source dataset fixed. This gives a clear statistical picture of the behaviour of the randomised algorithm. An important result is a conditional central limit theorem for the sketched dataset

that connects the Hadamard and Clarkson-Woodruff projections to the Gaussian sketch. The regularity conditions have a intuitive interpretation in terms of the geometry of the source dataset. We then analyse a large genetic dataset to compare the performance of different sketching algorithms and to test the asymptotic theory that we have developed.

4.2 Background and related work

Before proceeding, it is worth mentioning alternatives to sketching, in particular iterative methods for calculating the least squares coefficients $\boldsymbol{\beta}_F$. These include coordinate descent or stochastic gradient methods. Iterative methods are guaranteed to converge to $\boldsymbol{\beta}_F$ under very mild conditions. These iterative techniques assume that the entire dataset can stored in memory in a single location, or require regular communication if the full dataset is distributed across multiple sites. Sketching algorithms are not burdened by these memory and communication costs, with the drawback of no convergence guarantees to $\boldsymbol{\beta}_F$. Connections to iterative methods are postponed until the discussion, the focus for now is on the single pass estimators $\boldsymbol{\beta}_S$ and $\boldsymbol{\beta}_P$.

The purpose of this section is to review the existing theoretical framework for sketching algorithms. Sketching algorithms are largely motivated through worst case guarantees. We recap how these bounds can be developed before studying the statistical properties of the sketched estimators.

It will be helpful to define a number of quantities related to the full dataset before moving on. Let $TSS_F = \boldsymbol{y}^\mathsf{T}\boldsymbol{y}, RSS_F = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_F\|_2^2, MSS_F = \|\boldsymbol{X}\boldsymbol{\beta}_F\|_2^2$ and $R_F^2 = MSS_F/TSS_F$. These terms summarise the goodness of fit of the model. The total, residual and model sum of squares are given by TSS_F, RSS_F and MSS_F respectively, with $TSS_F = MSS_F + RSS_F$. The proportion of variance explained by the model is given by R_F^2. These values will be important in characterising the behaviour of $\boldsymbol{\beta}_S$ and $\boldsymbol{\beta}_P$.

4.2.1 Embedding bounds

A key concept in the construction of sketching algorithms is the notion of an ϵ-subspace embedding (Woodruff, 2014; Meng and Mahoney, 2013; Yang et al., 2015a).

Definition 4.1. *ϵ-subspace embedding.*
For a given $n \times d$ matrix $\boldsymbol{A}$, we call a $k \times n$ matrix $\boldsymbol{S}$ an ϵ-subspace embedding for $\boldsymbol{A}$, if for all vectors $\boldsymbol{z} \in \mathbb{R}^p$

$$(1-\epsilon)||\boldsymbol{A}\boldsymbol{z}||_2^2 \leq ||\boldsymbol{S}\boldsymbol{A}\boldsymbol{z}||_2^2 \leq (1+\epsilon)||\boldsymbol{A}\boldsymbol{z}||_2^2.$$

Speaking broadly, an ϵ-subspace preserves the linear structure of the columns of the original dataset up to some multiplicative $(1 \pm \epsilon)$ factor. In particular, if ϵ is small, the linear mapping $\boldsymbol{S}$ approximately preserves the covariance structure of the source dataset. Most theoretical arguments for sketching algorithms are predicated on the idea that the sketching matrix $\boldsymbol{S}$ is an ϵ-subspace embedding for the source dataset. The general notion is that it is possible to use a linear mapping $\boldsymbol{S}$ that reduces the sample size from n to k whilst preserving much of the linear information in the full dataset.

The issue of how to generate ϵ-subspace embeddings is deferred until section 3.2, the present focus will be on the utility of ϵ-subspace embeddings for linear regression problems. For now, assume that we have some method for generating ϵ-subspace embeddings for the source data matrix $\boldsymbol{A}$. It will be convenient to refer to $\widetilde{\boldsymbol{A}} = \boldsymbol{S}\boldsymbol{A}$ as an ϵ-subspace embedding of $\boldsymbol{A}$ if $\boldsymbol{S}$ is an ϵ-subspace embedding for $\boldsymbol{A}$. As regression is the focus from this point forward, we will define the source data matrix as $\boldsymbol{A} = [\boldsymbol{y}, \boldsymbol{X}]$, the sketched data matrix as $\widetilde{\boldsymbol{A}} = [\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}]$ and set $d = p + 1$.

The complete sketched estimator $\boldsymbol{\beta}_S$ is given by the least squares coefficients using the sketched responses $\widetilde{\boldsymbol{y}}$ and the sketched predictors $\widetilde{\boldsymbol{X}}$,

$$\boldsymbol{\beta}_S = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \|\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}\|_2^2.$$

An ϵ-subspace embedding is useful as it relates the sketched optimisation problem to the full dataset optimisation problem. If $\widetilde{\boldsymbol{A}} = [\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}]$ is an ϵ-subspace embedding of $\boldsymbol{A} = [\boldsymbol{y}, \boldsymbol{X}]$, it must hold that for all $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$(1-\epsilon)||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \leq ||\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}||_2^2 \leq (1+\epsilon)||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2.$$

If ϵ is small, minimising the sum of squared residuals on the sketched dataset is similar to minimising the sum of squared residuals on the full dataset. If this is the case, it can be expected that $\boldsymbol{\beta}_S$ will be close to $\boldsymbol{\beta}_F$. It is possible to establish the concrete bounds, that if $\widetilde{\boldsymbol{A}}$ is an ϵ-subspace embedding of $\widetilde{\boldsymbol{A}}$ (Sarlos, 2006),

$$\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_F\|_2^2 \leq \frac{\epsilon^2}{\sigma_{\min}^2(\boldsymbol{X})} RSS_F, \tag{4.3}$$

where $\sigma_{\min}(\boldsymbol{X})$ represents the smallest singular value of the design matrix $\boldsymbol{X}$. A very similar argument can be used to motivate the partial sketched estimator $\boldsymbol{\beta}_P$. Existing bounds for the partial sketch focus on the prediction error $\|\boldsymbol{X}\boldsymbol{\beta}_P - \boldsymbol{X}\boldsymbol{\beta}_F\|_2^2$ (Becker et al., 2015; Pilanci and Wainwright, 2016). To make a direct comparison to (4.3) we establish a bound on the coefficient error

Theorem 4.1. *Suppose that $\widetilde{\boldsymbol{X}}$ is an ϵ-subspace embedding of $\boldsymbol{X}$ with $\epsilon < 0.5$. Then the following bound holds,*

$$\|\boldsymbol{\beta}_P - \boldsymbol{\beta}_F\|_2^2 \leq \frac{4\epsilon^2}{\sigma_{\min}^2(\boldsymbol{X})} MSS_F \tag{4.4}$$

For proof see Chapter 5. The mild requirement that $\epsilon < 0.5$ is imposed so that the bound matches the functional form of the complete sketching bound (4.3). Comparing the partial sketching bound to (4.3), we see that the tightness of the bound is controlled by the model sum of squares as opposed to the residual sum of squares. The sensitivity of partial sketching to the model sum of squares as opposed to the residual sum of squares has been noted in previous on partial sketching (Dhillon et al., 2013; Pilanci and Wainwright, 2016; Becker et al., 2015). This suggests that the signal to noise ratio in the source dataset will be important when selecting which sketched estimator to use. A naive conclusion is that complete sketching is preferred when $RSS_F < 4MSS_F$, or equivalently $R_F^2 > 0.25$. Such a result is hardly prescriptive, as the worst case bound is not necessarily indicative of expected performance. A second point of interest is that if the $k \times n$ matrix $\boldsymbol{S}$ is an ϵ-subspace embedding for $\boldsymbol{A} = [\boldsymbol{y}, \boldsymbol{X}]$, it is also an ϵ-subspace embedding for $\boldsymbol{X}$. This suggests that it is reasonable to compute both $\boldsymbol{\beta}_P$ and $\boldsymbol{\beta}_S$ from a single sketch, although it is not clear how to combine the estimators into a single point estimator. These issues will be explored in more depth by examining the statistical properties of both complete and partial sketching. Before moving on to the statistical analysis we review some of the existing methods for generating ϵ-subspace embeddings.

4.2.2 Sketches

There are two general categories of distributions for the random matrix $\boldsymbol{S}$, data aware random projections and data oblivious random projections. A data aware random projection uses information in the source data $\boldsymbol{y}$, $\boldsymbol{X}$ to generate $\boldsymbol{S}$. In contrast, a data oblivious random projection can be sampled without knowledge of $\boldsymbol{y}$ or $\boldsymbol{X}$. Data oblivious projections are designed to produce ϵ-subspace embeddings for an arbitrary source data matrix with high probability. Our focus is on data oblivious random projections.

The Gaussian sketch was one of the first projections proposed for sketched regression (Sarlos, 2006). Recall that a Gaussian sketch is formed by independently sampling each element of $\boldsymbol{S}$ from a $N(0, 1/k)$ distribution. The drawback of the Gaussian sketch is that computation of the sketched data is quite demanding, taking $O(npk)$ operations. As such, there has been work on designing more computationally efficient random projections. The Hadamard sketch and the Clarkson-Woodruff sketch are two examples of more efficient methods for generating ϵ-subspace embeddings.

The Hadamard sketch is a structured random matrix (Ailon and Chazelle, 2009). The sketching matrix is formed as $\boldsymbol{S} = \Phi \boldsymbol{H}\boldsymbol{D}/\sqrt{k}$, where Φ is a $k \times n$ matrix and $\boldsymbol{H}$ and $\boldsymbol{D}$ are both $n \times n$ matrices. The fixed matrix $\boldsymbol{H}$ is a Hadamard matrix of order n. A Hadamard matrix is a square matrix with elements that are either $+1$ or -1 and orthogonal rows. Hadamard matrices do not exist for all integers n, the source dataset can be padded with zeroes so that a conformable Hadamard matrix is available. The random matrix $\boldsymbol{D}$ is a diagonal matrix where each nonzero element is an independent Rademacher random variable. The random matrix Φ subsamples k rows of $\boldsymbol{H}$ with replacement. The structure of the Hadamard sketch allows for fast matrix multiplication, reducing calculation of the sketched dataset to $O(nd \log k)$ operations.

The Clarkson-Woodruff sketch is a sparse random matrix (Clarkson and Woodruff, 2013). The projection can be represented as the product of two independent random matrices, $\boldsymbol{S} = \boldsymbol{\Gamma}\boldsymbol{D}$, where $\boldsymbol{\Gamma}$ is a random $k \times n$ matrix and $\boldsymbol{D}$ is a random $n \times n$ matrix. The matrix $\boldsymbol{\Gamma}$ is formed by choosing one element in each column independently and setting the entry to $+1$. The matrix $\boldsymbol{D}$ is a diagonal matrix where each nonzero element is an independent Rademacher random variable. This results in a sparse $\boldsymbol{S}$, where there is only one nonzero entry per column. The sparsity of the Clarkson-Woodruff sketch speeds up matrix multiplication, dropping the complexity of generating the sketched dataset to $O(nd)$.

4.2.3 Sketching examples

As examples, we demonstrate the construction of a Hadamard sketch and a Clarkson-Woodruff sketch, for $k = 3$, $n = 4$.

The Hadamard sketch matrix is formed as $\boldsymbol{S} = \Phi \boldsymbol{H}\boldsymbol{D}/\sqrt{k}$, where Φ is a $k \times n$ matrix and $\boldsymbol{H}$ and $\boldsymbol{D}$ are both $n \times n$ matrices. The fixed matrix $\boldsymbol{H}$ is a Hadamard matrix of order n. The random matrix $\boldsymbol{D}$ is a diagonal matrix where each nonzero element is an independent Rademacher random variable. The random matrix Φ subsamples k rows of $\boldsymbol{H}$ with replacement. The display below shows an example of the random projection. The first matrix in the display represents $\boldsymbol{\Phi H}$, a subsample of three rows from a 4×4 Hadamard matrix. In step 2, the diagonal matrix $\boldsymbol{D}$ is generated, with random Rademacher random variables along the diagonal. The diagonal elements are shown above the matrix. In step 3 the matrix multiplication $\boldsymbol{\Phi H D}$ is performed. This outputs the sketching matrix $\boldsymbol{S}$.

$$\begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{pmatrix} \underset{\text{step 2}}{\rightarrow} \begin{array}{c} \begin{matrix} +1 & -1 & +1 & +1 \\ D_{11} & D_{22} & D_{33} & D_{44} \end{matrix} \\ \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{pmatrix} \end{array} \underset{\text{step 3}}{\rightarrow} \begin{array}{c} \begin{matrix} +1 & -1 & +1 & +1 \\ \times & \times & \times & \times \end{matrix} \\ \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{pmatrix} \end{array} \underset{\text{output}}{\rightarrow} \begin{pmatrix} 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{pmatrix}$$

The Clarkson-Woodruff sketch is a sparse random matrix. The projection can be represented as the product of two independent random matrices, $\boldsymbol{S} = \boldsymbol{\Gamma}\boldsymbol{D}$, where $\boldsymbol{\Gamma}$ is a random $k \times n$ matrix and $\boldsymbol{D}$ is a random $n \times n$ matrix. The matrix $\boldsymbol{\Gamma}$ is formed by choosing one element in each column independently and setting the entry to $+1$. The matrix $\boldsymbol{D}$ is a diagonal matrix where each nonzero element is an independent Rademacher random variable. This results in a sparse $\boldsymbol{S}$, where there is only one nonzero entry per column. The display below shows an example of the random projection. The first matrix in the display represents $\boldsymbol{\Gamma}$, a random matrix where a single element in each column is set to one. In step 2, the diagonal

matrix $\boldsymbol{D}$ is generated, with random Rademacher random variables along the diagonal. The diagonal elements are shown above the matrix. In step 3 the matrix multiplication $\boldsymbol{\Gamma D}$ is performed. This outputs the sketching matrix $\boldsymbol{S}$.

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \underset{\text{step 2}}{\rightarrow} \begin{matrix} -1 & +1 & -1 & +1 \\ D_{11} & D_{22} & D_{33} & D_{44} \end{matrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \underset{\text{step 3}}{\rightarrow} \begin{matrix} +1 & -1 & +1 & +1 \\ \times & \times & \times & \times \end{matrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \underset{\text{output}}{\rightarrow} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & -1 & 1 & 0 \end{pmatrix}$$

Figure 4.1 shows examples of the three sketches for $k = 32, n = 36$.

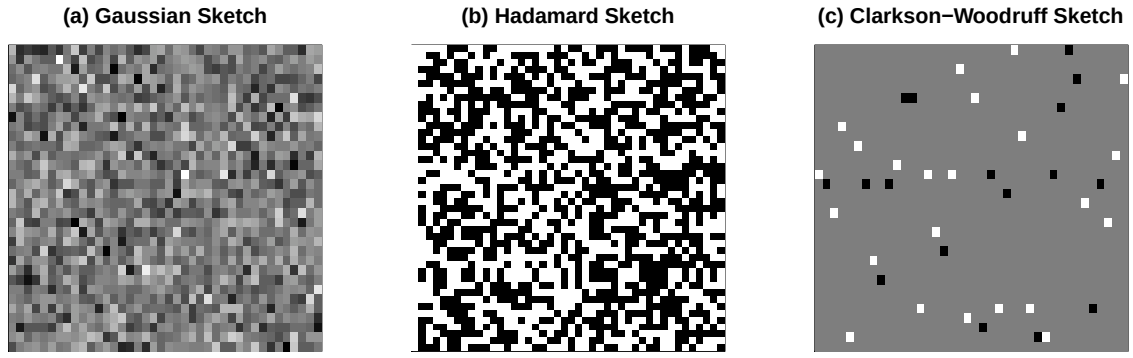


Figure 4.1: Sampled sketching matrices $\boldsymbol{S}$ for $k = 32, n = 36$. Elements in the sketching matrix are coloured based on the value. One and negative one are coloured as black and white respectively. Intermediate values are in shades of grey.

4.2.4 Sketching bounds

Data oblivious sketches are designed to give an ϵ-subspace embedding for an arbitrary source dataset with at least probability $(1 - \delta)$. Sketching algorithms are appealing for large n problems as the required k to attain the (δ, ϵ) bound is independent of n for the Gaussian and Clarkson-Woodruff sketches, and very weakly dependent on n for the Hadamard sketch. Table 4.1 summarises existing results on the necessary k to attain the (ϵ, δ) bound. Probabilistic worst case bounds for sketched regression are formed by noting that if a sketch produces an ϵ-subspace embedding with probability at least $(1 - \delta)$, then the bounds in section 2 must hold with probability at least $(1 - \delta)$. Woodruff (2014) gives an excellent survey of work in this area. We consider embedding probabilities in more detail in Chapter 6.

As mentioned, data aware random projections can also be used to generate ϵ-subspace embeddings. Existing data aware projections perform weighted sampling with replacement from the source dataset. As such, data aware sketching methods are closely related to resampling methods such as the bootstrap and the jacknife (Ma and Sun, 2015). We focus on data oblivious random projections, where there is no direct connection to resampling methods. The Gaussian sketch is mathematically tractable, and it is possible to establish a number of exact finite sample results regarding the performance of the sketched estimators. In the next section, we obtain the exact distribution of $\boldsymbol{\beta}_S$ and the bias and variance of $\boldsymbol{\beta}_P$. This provides guidance on issues regarding the relative efficiency of complete to partial sketching.

Sketch	Sketching time	Required sketch size k
Gaussian	$O(ndk)$	$O((d+\log(1/\delta))/\epsilon^2)$
Hadamard	$O(nd\log k)$	$O((\sqrt{d}+\sqrt{\log n})^2(\log(d/\delta))/\epsilon^2)$
Clarkson-Woodruff	$O(nd)$	$O(d^2/\delta\epsilon^2)$

Table 4.1: Properties of different data oblivious random projections (Woodruff, 2014). The third column refers to the necessary sketch size k to obtain an ϵ-subspace embedding for an arbitrary $n \times d$ source dataset with at least probability $(1-\delta)$.

4.3 Gaussian sketching

4.3.1 Complete sketching

The Gaussian sketch is mathematically tractable, and it is possible to establish a number of exact finite sample results regarding the performance of the sketched estimators. In this section we will develop the distribution of $\boldsymbol{\beta}_S$ when using a Gaussian sketch. As mentioned previously, all results treat $\boldsymbol{y}$ and $\boldsymbol{X}$ as fixed. The variability in $\boldsymbol{\beta}_S$ is solely due to the use of the random sketching matrix $\boldsymbol{S}$. Let $(\widetilde{y}_j, \widetilde{\boldsymbol{x}}_j^\mathsf{T})^\mathsf{T}$ refer to the jth row in the sketched data matrix $\widetilde{\boldsymbol{A}} = [\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}]$ for $j = 1, \ldots, k$. Similarly, let $\boldsymbol{s}_j^\mathsf{T}$ denote the jth row in the sketching matrix $\boldsymbol{S}$. The sketched dataset consists of k random units $(\widetilde{y}_j, \widetilde{\boldsymbol{x}}_j^\mathsf{T})$ for $j = 1, \ldots, k$. The jth sketched response is given by $\widetilde{y}_j = \boldsymbol{s}_j^\mathsf{T}\boldsymbol{y}$, and the jth sketched predictor is calculated as $\widetilde{\boldsymbol{x}}_j = \boldsymbol{s}_j^\mathsf{T}\boldsymbol{X}$ for $j = 1, \ldots, k$. The k sketched instances are independently distributed, because rows of the sketching matrix are independent.

We take an indirect route to find the distribution of $\boldsymbol{\beta}_S$, by focusing on the distribution of the sketched data $\widetilde{\boldsymbol{A}} = [\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}]$ conditional on the original dataset $\boldsymbol{A} = [\boldsymbol{y}, \boldsymbol{X}]$. The initial step is to decompose the joint distribution on the sketched responses and predictors as the product of a marginal and conditional distribution. Specifically,

$$p(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}} | \boldsymbol{y}, \boldsymbol{X}) = p(\widetilde{\boldsymbol{y}} | \widetilde{\boldsymbol{X}}, \boldsymbol{y}, \boldsymbol{X}) p(\widetilde{\boldsymbol{X}} | \boldsymbol{y}, \boldsymbol{X}).$$

It can be shown that $p(\widetilde{\boldsymbol{y}} | \widetilde{\boldsymbol{X}}, \boldsymbol{y}, \boldsymbol{X}) p(\widetilde{\boldsymbol{X}} | \boldsymbol{y}, \boldsymbol{X})$ has the structure of a hierarchical Gaussian linear model. We first show that the sketched dataset has a multivariate normal distribution, conditional on the source dataset. This follows as the sketched dataset can be expressed as a linear combination of Gaussian random variables. Specifically, row j in the sketched dataset is given by $\widetilde{\boldsymbol{a}}_j = (\widetilde{y}_j, \widetilde{\boldsymbol{x}}_j) = \boldsymbol{s}_j \boldsymbol{A}$. To be clear, it is helpful to express $(\widetilde{y}_j, \widetilde{\boldsymbol{x}}_j)$ as a column vector

$$\begin{bmatrix} \widetilde{y}_j \\ \widetilde{\boldsymbol{x}}_j^\mathsf{T} \end{bmatrix} = \boldsymbol{A}^\mathsf{T} \boldsymbol{s}_j^\mathsf{T}.$$

Conditional on $\boldsymbol{A} = [\boldsymbol{y}, \boldsymbol{X}]$, $\boldsymbol{A}^\mathsf{T}\boldsymbol{s}_j^\mathsf{T}$ is a linear combination of independent Gaussians as $\boldsymbol{s}_j^\mathsf{T} \sim N(\boldsymbol{0}, \boldsymbol{I}_d/k)$. As affine transformations of Gaussians are also multivariate normal, $(\widetilde{y}_j, \widetilde{\boldsymbol{x}}_j)$ must then be jointly normally distributed, conditional on the source data $(\boldsymbol{y}, \boldsymbol{X})$. It is easily shown that the joint distribution of the sketched responses and predictors is then

$$\begin{bmatrix} \widetilde{y}_j \\ \widetilde{\boldsymbol{x}}_j^\mathsf{T} \end{bmatrix} \Bigg| \, \boldsymbol{y}, \boldsymbol{X} \sim N\left(\begin{bmatrix} 0 \\ \boldsymbol{0} \end{bmatrix}, \frac{1}{k} \begin{bmatrix} \boldsymbol{y}^\mathsf{T}\boldsymbol{y} & \boldsymbol{y}^\mathsf{T}\boldsymbol{X} \\ \boldsymbol{X}^\mathsf{T}\boldsymbol{y} & \boldsymbol{X}^\mathsf{T}\boldsymbol{X} \end{bmatrix} \right), \quad \text{independently for } j = 1, \ldots, k.$$

Standard results on multivariate normals give that the conditional distribution of $\widetilde{y}_j$ given $\widetilde{\boldsymbol{x}}_j$ is also normal. A routine calculation shows that the conditional mean is related to $\boldsymbol{\beta}_F$, that is $E_S[\widetilde{y}_j \mid \widetilde{\boldsymbol{x}}_j, \boldsymbol{y}, \boldsymbol{X}] = \widetilde{\boldsymbol{x}}_j \boldsymbol{\beta}_F$. The subscript S is used on the expectation operator to emphasise that only random quantity is the sketching matrix. The conditional variance is related to the prediction error on the source dataset

RSS_F,

$$\begin{aligned}\text{var}_S\left(\widetilde{y}_j|\widetilde{\boldsymbol{x}}_j, \boldsymbol{y}, \boldsymbol{X}\right) &= \frac{1}{k}\left[\boldsymbol{y}^\mathsf{T}\boldsymbol{y} - \boldsymbol{y}^\mathsf{T}\boldsymbol{X}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}\right] \\ &= \frac{1}{k}RSS_F.\end{aligned}$$

The subscript S is again used to recognise that the source of the variance is the random sketching matrix, the source dataset is fixed. The step in the second line follows from sum of squares partitions in linear models (Searle, 1997, Chapter 3). Therefore, the conditional distribution of $\widetilde{y}_j$ given the sketched predictors $\widetilde{\boldsymbol{x}}_j$ and the source dataset $\boldsymbol{y}, \boldsymbol{X}$ is

$$\widetilde{y}_j \mid \widetilde{\boldsymbol{x}}_j, \boldsymbol{y}, \boldsymbol{X} \sim N\left(\widetilde{\boldsymbol{x}}_j\boldsymbol{\beta}_F, \frac{RSS_F}{k}\right) \quad \text{independently for } j = 1, \ \ldots, \ k.$$

This is the exact form of a standard Gaussian linear model. The distribution $p(\widetilde{\boldsymbol{X}}|\boldsymbol{y}, \boldsymbol{X})$ is easily obtained as the marginal distribution of $\widetilde{\boldsymbol{x}}_j$ is also multivariate normal,

$$\widetilde{\boldsymbol{x}}_j^\mathsf{T} \sim N\left(\boldsymbol{0}, \boldsymbol{X}^\mathsf{T}\boldsymbol{X}/k\right), \quad \text{independently for } j = 1, \ \ldots, \ k.$$

The sketching process can be described using the following hierarchical model,

$$\begin{aligned}\widetilde{\boldsymbol{y}} \mid \widetilde{\boldsymbol{X}}, \boldsymbol{y}, \boldsymbol{X} &\sim N\left(\widetilde{\boldsymbol{X}}\boldsymbol{\beta}_F, \frac{RSS_F}{k}\boldsymbol{I}_k\right), \\ \widetilde{\boldsymbol{X}} \mid \boldsymbol{y}, \boldsymbol{X} &\sim MN\left(\boldsymbol{0}_{k\times p}, \boldsymbol{I}_k, \frac{1}{k}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\right).\end{aligned}$$

A Gaussian sketch effectively simulates a series of observations from a Gaussian linear model parametrised in terms of $\boldsymbol{\beta}_F$ and σ_F^2, where the design matrix has a matrix normal distribution. We now turn to the distribution of $\boldsymbol{\beta}_S$. The distribution of $\boldsymbol{\beta}_S$ conditional on the sketched predictors follows immediately from standard results on linear models (Searle, 1997, Chapter 3).

$$[\boldsymbol{\beta}_S \mid \widetilde{\boldsymbol{X}}, \boldsymbol{y}, \boldsymbol{X}] \sim N\left(\boldsymbol{\beta}_F, \frac{RSS_F}{k}(\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})^{-1}\right). \tag{4.5}$$

To obtain the marginal distribution of $\boldsymbol{\beta}_S$ it is necessary to integrate over the random sketched design matrix $\widetilde{\boldsymbol{X}}$. From properties of the normal distribution (Eaton, 2007), it is possible to show $(\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}}) \mid \boldsymbol{y}, \boldsymbol{X} \sim \text{Wishart}(k, \boldsymbol{X}^\mathsf{T}\boldsymbol{X}/k)$. As such,

$$(\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})^{-1} \mid \boldsymbol{y}, \boldsymbol{X} \sim \text{InvWishart}\left(k, k(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\right).$$

As seen in equation (4.5), $\boldsymbol{\beta}_S$ is normally distributed when conditioned on the random Inverse-Wishart matrix $(\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})^{-1}$. The marginal distribution of $\boldsymbol{\beta}_S$ can then be described using the Normal Inverse-Wishart distribution (Gelman et al., 2014, p.73). The following theorem characterises the distribution of $\boldsymbol{\beta}_S$ under the Gaussian sketch.

Theorem 4.2. *Suppose $\boldsymbol{\beta}_S$ is computed using a Gaussian sketch and $k > p + 1$. The conditional distribution of $\boldsymbol{\beta}_S$ is*

$$(i) \ \boldsymbol{\beta}_S|\widetilde{\boldsymbol{X}}, \boldsymbol{y}, \boldsymbol{X} \sim N\left(\boldsymbol{\beta}_F, \frac{n\sigma_F^2}{k}\left(\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}}\right)^{-1}\right).$$

The marginal distribution of $\boldsymbol{\beta}_S$ is

$$(ii) \ \boldsymbol{\beta}_S|\boldsymbol{y}, \boldsymbol{X} \sim \text{Student}\left(\boldsymbol{\beta}_F, \frac{n\sigma_F^2}{k-p+1}\left(\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\right)^{-1}, \ k-p+1\right).$$

For the proof see Chapter 5.

An immediate application of result (i) is the ability to generate exact confidence intervals for the elements of $\boldsymbol{\beta}_S$, methodology that does not appear to be present in the existing literature. It is also

possible to estimate exact joint confidence regions for the entire vector $\boldsymbol{\beta}_S$. Again assuming that $k > p+1$, it should be noted that the variance of $\boldsymbol{\beta}_S$,

$$\text{var}(\boldsymbol{\beta}_S|\boldsymbol{y}, \boldsymbol{X}) = \frac{RSS_F}{(k-p+1)}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1} \tag{4.6}$$

is not dependent on the compression ratio k/n. Although RSS_F can be expected to grow linearly with n, this will generally be counterbalanced by $(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}$ decreasing linearly with n. The distribution of the approximation error $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_F\|_2$ will largely be controlled by the target dimension k. This speaks to the defining characteristic of sketching algorithms, that given a fixed k, the stochastic approximation error does not necessarily increase with size of the original dataset n. Probabilistic worst case bounds on the error $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_F\|_2$ can also be obtained by making an appeal to Chebyschev's inequality.

4.3.2 Partial sketching

Partial sketching was first proposed by Dhillon et al. (2013) using uniform subsampling, and later studied for general sketches by Pilanci and Wainwright (2016). Existing results on partial sketching highlight that the model sum of squares influences the approximation error of the partial sketched estimator $\boldsymbol{\beta}_P$. An important finding was that the variance of the complete sketching estimator is dependent on the residual sum of squares. It is simple to see that the variance of the partial sketched estimator will not be a function of the residual sum of squares. From the normal equations it holds that $\boldsymbol{X}^\mathsf{T}\boldsymbol{y} = \boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta}_F$. Using this property, we see that conditional on $\boldsymbol{y}, \boldsymbol{X}$, the variance of the random linear combination $\boldsymbol{\beta}_P = (\boldsymbol{X}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta}_F$ will be a function of the covariates and the fitted values. The residual vector has no influence on the variance of the partial sketching estimator, and as such the variance of $\boldsymbol{\beta}_P$ will not be related to the residual sum of squares. This suggests that when the noise level is high, partial sketching may become preferrable to complete sketching. This idea has been touched on in the existing literature, but specific guidelines are lacking (Becker et al., 2015; Dhillon et al., 2013). A statistical analysis can provide some insight into this issue.

The hierarchical model for complete sketching gave an intuitive statistical perspective on the mechanics of the algorithm. Partial sketching seems to lack a similar conceptual device. The least squares coefficients can be represented as the solution to the linear system of the equations $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^\mathsf{T}\boldsymbol{y}$. Partial sketching simply returns the solution, $\boldsymbol{b}$, to the approximate linear system $\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}}\boldsymbol{b} = \boldsymbol{X}^\mathsf{T}\boldsymbol{y}$. Lacking a convenient representation for the estimator, we must proceed in a more pedestrian manner. The mean square error of the estimator $\boldsymbol{\beta}_P$ can be determined using only mean and variance information, and this will be the goal for now. The key observation is that $(\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})^{-1}|\boldsymbol{y}, \boldsymbol{X} \sim \text{InvWishart}\left(k, k(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\right)$. Conditional on $\boldsymbol{y}, \boldsymbol{X}$, the estimator $\boldsymbol{\beta}_P = (\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$ is a linear combination of the elements of an Inverse-Wishart random variable. However, this is a non-standard distribution and it is difficult to directly express the distribution function of $\boldsymbol{\beta}_P$. Despite this, it is straightforward to determine the mean and variance of $\boldsymbol{\beta}_P$. From properties of the Inverse-Wishart distribution, it can be seen that the partial sketched estimator is biased, with mean

$$E_S[\boldsymbol{\beta}_P|\boldsymbol{y}, \boldsymbol{X}] = \frac{k}{(k-p-1)}\boldsymbol{\beta}_F,$$

where it is assumed that $k > p+3$. This motivates an alternative unbiased estimator

$$\boldsymbol{\beta}_P^* = \frac{(k-p-1)}{k}(\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}.$$

Determining the variance of $\boldsymbol{\beta}_P$ and the unbiased $\boldsymbol{\beta}_P^*$ is a more lengthy computation (see Chapter 5). Skipping the work, the variance of the biased estimator $\boldsymbol{\beta}_P$ is

$$\text{var}(\boldsymbol{\beta}_P|\boldsymbol{y}, \boldsymbol{X}) = \frac{k^2}{(k-p)(k-p-1)(k-p-3)}\left(MSS_F(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1} + \frac{(k-p+1)}{(k-p-1)}\boldsymbol{\beta}_F\boldsymbol{\beta}_F^\mathsf{T}\right). \tag{4.7}$$

The variance of the unbiased estimator $\boldsymbol{\beta}_P^*$ is

$$\text{var}(\boldsymbol{\beta}_P^*|\boldsymbol{y}, \boldsymbol{X}) = \frac{(k-p-1)}{(k-p)(k-p-3)} \left(MSS_F(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1} + \frac{(k-p+1)}{(k-p-1)}\boldsymbol{\beta}_F\boldsymbol{\beta}_F^\mathsf{T} \right). \tag{4.8}$$

The variances of $\boldsymbol{\beta}_P$ and $\boldsymbol{\beta}_P^*$ have a similar structure to the variance of $\boldsymbol{\beta}_S$. The main point of difference is that the variance of $\boldsymbol{\beta}_S$ depends on the residual sum of squares, whereas the variance of $\boldsymbol{\beta}_P$ and $\boldsymbol{\beta}_P^*$ depends on the model sum of squares.

As mentioned the explicit form of the sampling distribution is hard to obtain, but by making a connection with method of moments estimation it is possible to establish asymptotic normality of both $\boldsymbol{\beta}_P$ and $\boldsymbol{\beta}_P^*$ as k tends to infinity. This motivates the construction of approximate confidence intervals. As the exact variance is unknown we propose the following estimator

$$\widehat{\text{var}}(\boldsymbol{\beta}_P^*|\boldsymbol{y}, \boldsymbol{X}) = \frac{(k-p-1)}{(k-p)(k-p-3)} \left(\left(\frac{k-p-1}{k} \right) MSS_S(\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})^{-1} + \boldsymbol{\beta}_P^*\boldsymbol{\beta}_P^{*\mathsf{T}} \right). \tag{4.9}$$

4.3.3 Relative efficiency

The relative efficacy of complete and partial sketching is also of interest. As the plug in estimator $\boldsymbol{\beta}_P$ has a higher mean square error than $\boldsymbol{\beta}_P^*$, it will not be considered in this section. The performance of the complete sketching estimator $\boldsymbol{\beta}_S$ and the unbiased partial sketched estimator $\boldsymbol{\beta}_P^*$ will be compared in terms of mean squared error. As both $\boldsymbol{\beta}_F$ and $\boldsymbol{\beta}_P^*$ are unbiased, the mean squared error can be computed using their respective covariance matrices, that is

$$E_S\left(\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_F\|_2^2 \mid \boldsymbol{y}, \boldsymbol{X} \right) = \text{tr}(\text{var}(\boldsymbol{\beta}_S)),$$
$$E_S\left(\|\boldsymbol{\beta}_P^* - \boldsymbol{\beta}_F\|_2^2 \mid \boldsymbol{y}, \boldsymbol{X} \right) = \text{tr}(\text{var}(\boldsymbol{\beta}_P^*)).$$

Comparing (4.6) and (4.8), the variance of $\boldsymbol{\beta}_P^*$ is dependent on MSS_F, whereas the variance of $\boldsymbol{\beta}_S$ is dependent on RSS_F. This suggests that the signal to noise ratio in the source dataset will be an influential factor in determining which estimator is more efficient. When R_F^2 is close to one we expect complete sketching to be be orders of magnitude more efficient than partial sketching, and when R_F^2 is close to zero, we expect partial sketching to be orders of magnitude more efficient than complete sketching.

4.3.4 Combined estimator

So far we have assumed that an analyst much choose between one of the two methods. Obtaining both $\boldsymbol{\beta}_P^*$ and $\boldsymbol{\beta}_S$ from a single sketch is computationally cheap, and may be an attractive strategy. The most demanding operation with the sketched data is calculating $(\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})^{-1}$. Given this quantity it is economical to compute both $\boldsymbol{\beta}_S$ and $\boldsymbol{\beta}_P^*$. Becker et al. (2015) mention they are presently investigating such a strategy, but do not give any details. Our motivation for a combined estimator is driven by the fact even when using a single sketch $(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}})$, the two estimators are uncorrelated, that is $\text{cov}\left(\boldsymbol{\beta}_P^*, \boldsymbol{\beta}_S\right) = \boldsymbol{0}$. This is established by taking iterated expectations, and using the hierarchical model established in section 4.3.1 (see Chapter 5). A simple strategy is then to take a weighted combination of $\boldsymbol{\beta}_S$ and $\boldsymbol{\beta}_P^*$. A combined estimator $\boldsymbol{\beta}_C$ can be defined as

$$\boldsymbol{\beta}_C = \alpha\boldsymbol{\beta}_S + (1-\alpha)\boldsymbol{\beta}_P^*,$$

for some $0 < \alpha < 1$. The value of α that minimises the mean square error is

$$\alpha_{\text{opt}} = \frac{\text{tr}(\text{var}(\boldsymbol{\beta}_P^*))}{\text{tr}(\text{var}(\boldsymbol{\beta}_P^*)) + \text{tr}(\text{var}(\boldsymbol{\beta}_S))}.$$

The weight given to the $\boldsymbol{\beta}_S$ is related to the relative efficiency of the two estimators. Use of the weighted estimator is expected to be most beneficial when the signal to noise ratio is moderate, that is $R_F^2 \approx 0.5$. When the signal to noise ratio is either very high or very low, there is little gain from using the weighted estimator as either the complete or partial estimator will dominate.

4.4 Asymptotics

4.4.1 Preliminaries

Finite sample distributions of random projection estimators can be mathematically intractable, and as such asymptotic analysis can be a powerful tool (Li et al., 2006). It is a very difficult task to establish meaningful finite sample results for the Hadamard and Clarkson-Woodruff sketches, as they are discrete distributions over a very large combinatorial space. The explicit finite sample distribution of the sketched estimators can be written as a sum over all these possible combinations, but such a representation is not very informative. Instead, it is useful to study the large n distribution of the estimators $\boldsymbol{\beta}_S$ and $\boldsymbol{\beta}_P$ to obtain an interpretable expression.

As $\boldsymbol{\beta}_F$ is the estimand in sketching algorithms, this requires conditioning on the source data in the asymptotic analysis. We make no assumption on the nature of the data generating process. To elaborate, let $\boldsymbol{A}_{(n)} = [\boldsymbol{y}_{(n)}, \boldsymbol{X}_{(n)}]$ represent the $n \times d$ source data matrix of full column rank. Any source data matrix $\boldsymbol{A}_{(n)}$ has a set of associated least squares coefficients, which will here be denoted $\boldsymbol{\beta}_F^{(n)}$. The overall goal is to determine the asymptotic form of the distributions $p(\boldsymbol{\beta}_S|\boldsymbol{A}_{(n)})$ and $p(\boldsymbol{\beta}_P^*|\boldsymbol{A}_{(n)})$ for some arbitrary large dataset $\boldsymbol{A}_{(n)}$.

To take limits, we employ a fixed sequence of $n \times d$ datasets, all of rank d. In the regression scenario this amounts to assuming that $\boldsymbol{X}_{(n)}$ is of full column rank and that $\boldsymbol{y}_{(n)}$ is not a perfect linear combination of the columns of $\boldsymbol{X}_{(n)}$ for all n. Conditioning on $\boldsymbol{A}_{(n)}$ is effectively the same as treating the full dataset as an arbitrary sequence of constants A_{ij} for $i = 1, \ldots, n$, $j = 1, \ldots, d$. This is analogous to large sample results for regression models where the design matrix is treated as arbitrary set of constants, and the random variables of interest are the error terms, for example see Van Der Vaart (1998, section 2.5). Here the source dataset is treated as a sequence of constants and the random variables of interest are the elements of the sketching matrix.

The asymptotic analysis is carried out in two stages. The initial step is to establish asymptotic normality of the sketched dataset. This is then followed by an analysis of the limiting distribution of $\boldsymbol{\beta}_S$, and $\boldsymbol{\beta}_P^*$. There is some related work by Ma et al. (2015) who develop asymptotic expressions for the bias and variance of data aware sketched regression estimators, where limits are taken in the sketch size k. Our work is different as we study data oblivious random projections and take limits in n, which is perhaps more natural in the Big Data setting.

4.4.2 Sketching central limit theorem

Using a Gaussian sketch, treating the source data matrix $\boldsymbol{A}$ as fixed:

$$[\widetilde{\boldsymbol{A}}|\boldsymbol{A}] \sim MN(\boldsymbol{0}, \boldsymbol{I}_k, \boldsymbol{A}^\mathsf{T}\boldsymbol{A}/k). \tag{4.10}$$

Each row is statistically independent, and marginally normally distributed with covariance matrix $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}/k$. We would like to establish an analogous asymptotic result for the Clarkson-Woodruff and Hadamard sketches. In (4.10), the source dataset is treated as fixed. The original data matrix $\boldsymbol{A}$ can originate from any data generating process. The source dataset could arise from a spatial point process, a collection of time series or a Gaussian Markov random field. When we treat the source data matrix $\boldsymbol{A}$ as fixed we are completely agnostic as to the data generating process that led to the creation of $\boldsymbol{A}$. We would like to establish an asymptotic result in this vein, where we make minimal assumptions on $\boldsymbol{A}$. We have gone to considerable effort to establish a conditional central limit theorem, where the source dataset is treated as an arbitrary sequence of fixed constants. This is so our expedition into asymptopia remains tied as closely to reality as possible. In practice, the large dataset that we apply a random projection to will be a fixed file living on a server or hard drive. We would like to describe the distribution of the random sketched dataset in relation to the fixed source dataset. This is the distribution that is reported in (4.10) for the Gaussian sketch. We want to understand the asymptotic form of this conditional distribution for

the Hadamard and Clarkson-Woodruff projections. The only random variables in the sequence of distributions that we study are elements of the random projections. An unconditional central limit theorem for sparse sketching matrices with independent entries is given in Li et al. (2006). We cannot easily extend their method as we wish to establish a conditional central limit theorem and the Clarkson-Woodruff sketch and the Hadamard sketch have dependent entries. Our method of proof differs in many significant aspects.

Under some regularity conditions the Hadamard and Clarkson-Woodruff sketches produce sketched data that asymptotically has the same matrix normal distribution as under the Gaussian sketch. Although asymptotic normality may not be particularly surprising seeing as the sketched data are linear combinations of random vectors, the proof is not immediate due to the dependence in the Hadamard and Clarkson-Woodruff sketches and our conditional perspective. The difficulties we face are most easily illustrated for the Clarkson-Woodruff sketch.

Algorithm 4.1 Clarkson-Woodruff sketch

$\widetilde{\boldsymbol{A}} \leftarrow \boldsymbol{0}$ Initialise sketched dataset as $k \times d$ matrix of zeroes
For $i = 1$ to $i = n$
  Sample $z \sim \text{Uniform}(1, \ldots, k)$ Sample random index
  Sample $r \sim \text{Uniform}(-1, +1)$ Sample random sign
  $\widetilde{\boldsymbol{A}}_z \leftarrow r \times \boldsymbol{A}_i + \widetilde{\boldsymbol{A}}_z$ Multiply by r and add to row z in sketch
Output $\widetilde{\boldsymbol{A}}$ Output sketched dataset

The behaviour of the Clarkson-Woodruff sketch can be represented as a many to less mapping. Each row in the source dataset is assigned to a single row in the sketched dataset. The Clarkson-Woodruff sketch has an alternative streaming construction that highlights this property, given in Algorithm 4.1. As each row in the source dataset only contributes to a single row in the sketched dataset, it might be expected that this results in some statistical dependence amongst the rows of the sketched dataset. The concept of dependence here is again conditional on the source dataset $\boldsymbol{A}$ being known. The independence in (4.10) refers to the fact that knowledge of any particular row in the sketched dataset does not help to predict any other row in the sketched dataset, given that we know the full source dataset $\boldsymbol{A}$. This definition of independence has interesting implications for the Clarkson-Woodruff sketch. Suppose that we have the following source dataset where $n = 3$ and $d = 2$:

$$\boldsymbol{A} = \begin{bmatrix} 10 & 10 \\ 1 & -1 \\ 0.1 & 0.1 \end{bmatrix} \tag{4.11}$$

Suppose that we take a Clarkson-Woodruff sketch of size $k = 2$ to obtain the $k \times d$ matrix $\widetilde{\boldsymbol{A}}$. Suppose we reveal the first row to an outside observer who then has to predict the second row in the sketched data matrix. The observer knows

$$\widetilde{\boldsymbol{A}} = \begin{bmatrix} 11 & 9 \\ ? & ? \end{bmatrix} \tag{4.12}$$

Conditional on the observer having access to the source data matrix $\boldsymbol{A}$ (4.11) they could easily reason that the second row in $\widetilde{\boldsymbol{A}}$ will either be $(0.1, 0.1)$ or $(-0.1, -0.1)$. The observer can piece this together by realising that rows 1 and 2 in the source dataset must have been assigned to the first row in the sketched dataset. Knowledge of $\boldsymbol{A}$ and the first row in $\widetilde{\boldsymbol{A}}$ is enough to reverse engineer the sketched mapping and to predict the second row in the sketched data matrix. For asymptotic equivalence with the Gaussian sketch, we require no information gain from revealing any row of the sketched data matrix given that we have access to the full source dataset $\boldsymbol{A}$. The predictive ability of any row in the sketched data matrix needs

to dissipate with n even after accounting for the fact that we are always conditioning on knowing the full source dataset $\boldsymbol{A}$. We will see that we can obtain asymptotic independence of the sketched dataset rows conditional on knowing the full source dataset under mild regularity conditions. Additionally, although it seems each row in the sketched dataset will be marginally normally distributed, it is not clear if joint asymptotic normality over all rows will hold. Similar conundrums arise when examining the Hadamard sketch in detail.

The $k \times d$ random matrix $\widetilde{\boldsymbol{A}}$ is the output of a stochastic process governed by the fixed $n \times d$ source dataset $\boldsymbol{A}_{(n)}$ and the distribution of the random $k \times n$ sketching matrix $\boldsymbol{S}$. The sketched dataset is a linear combination of random vectors, the number of which increases with n. As such, we can expect $\widetilde{\boldsymbol{A}}$ to demonstrate some stable limiting behaviour as n grows larger. Under an assumption on the limiting leverage scores of the source data matrix, we can establish a central limit theorem for the sketched dataset. Recall the singular value decomposition of the source dataset $\boldsymbol{A}_{(n)} = \boldsymbol{U}_{(n)}\boldsymbol{D}_{(n)}\boldsymbol{V}_{(n)}^{\mathsf{T}}$. The leverage scores for observation i in the source dataset is defined as $\|\boldsymbol{u}_{(n)i}\|_2^2$ where $\boldsymbol{u}_{(n)i}^{\mathsf{T}}$ gives row i in $\boldsymbol{U}_{(n)}$. The leverage scores of the observations in the source data matrix have been identified an important structural property of sketching algorithms (Mahoney and Drineas, 2016). Assumption 1 highlights their role in establishing asymptotic normality of the sketched data matrix.

Assumption 1 Let the singular value decomposition of the $n \times d$ source dataset be given by $\boldsymbol{A}_{(n)} = \boldsymbol{U}_{(n)}\boldsymbol{D}_{(n)}\boldsymbol{V}_{(n)}^{\mathsf{T}}$. Let $\boldsymbol{u}_{(n)i}^{\mathsf{T}}$ give the ith row in $\boldsymbol{U}_{(n)}$. Assume that the maximum leverage score tends to zero, that is

$$\lim_{n\to\infty} \max_{i=1,\ldots,n} \|\boldsymbol{u}_{(n)i}\|_2^2 = 0.$$

Theorem 4.3 gives the sketching central limit theorem.

Theorem 4.3. *Consider a fixed sequence of arbitrary $n \times d$ data matrices $\boldsymbol{A}_{(n)}$, where d is fixed. Let $\boldsymbol{A}_{(n)} = \boldsymbol{U}_{(n)}\boldsymbol{D}_{(n)}\boldsymbol{V}_{(n)}^{\mathsf{T}}$ represent the singular value decomposition of $\boldsymbol{A}_{(n)}$. Let $\boldsymbol{S}$ be a $k \times n$ Hadamard or Clarkson-Woodruff sketching matrix where k is also fixed. Suppose that Assumption 1 on the maximum leverage score is satisfied. Then as n tends to infinity with k and d fixed,*

$$[\widetilde{\boldsymbol{A}}\boldsymbol{V}_{(n)}\boldsymbol{D}_{(n)}^{-1} \mid \boldsymbol{A}_{(n)}] \xrightarrow{d} \mathrm{MN}(\boldsymbol{0}, \boldsymbol{I}_k, \boldsymbol{I}_d/k).$$

The proof of Theorem 4.3 is given in Chapter 5. Heuristically, for large n we expect the matrix normal result (4.10) to approximately hold for the Hadamard and Clarkson-Woodruff sketches. The significance of Assumption 1 is perhaps best explained by making a connection to a version of the Lindeberg-Feller theorem for triangular arrays of uniformly bounded random variables (Billingsley, 1999).

Theorem 4.4 (Billingsley, 1999)**.** *For each $n \in \mathbb{N}$, let $Z_{n1}, Z_{n2}, \ldots, Z_{nr_n}$ be a sequence of independent random variables with $\mathbb{E}(Z_{ni}) = 0$ and $\mathrm{var}(Z_{ni}) = \sigma_{ni}^2$ for $i = 1, \ldots, r_n$. Let $s_n^2 = \sum_{i=1}^{r_n} \sigma_{ni}^2$ and assume that $r_n \to \infty$ as $n \to \infty$. Suppose that we can form a sequence of upper bounds $(K_n)_{n\in\mathbb{N}}$ such that*

$$|Z_{ni}| \leq K_n \text{ almost surely for } i = 1, \ldots, r_n.$$

Then if $K_n/s_n \to 0$ as $n \to \infty$ we have the convergence in distribution

$$\frac{1}{s_n}\sum_{i=1}^{r_n} Z_{ni} \xrightarrow{d} N(0,1)$$

In Theorem 4.4 the condition that $K_n/s_n \to 0$ ensures that no random variable in a particular row of the array has to much pull over the sum $\sum_{i=1}^{r_n} Z_{ni}$. A triangular array of random variables satisfying the conditions in Theorem 4.4 is often said to be uniformly asymptotically negligible in that no single term has undue influence over the random sum. We can make an analogy to the leverage score condition in the sketching central limit theorem (Theorem 4.3). The sum of the statistical leverage scores is always

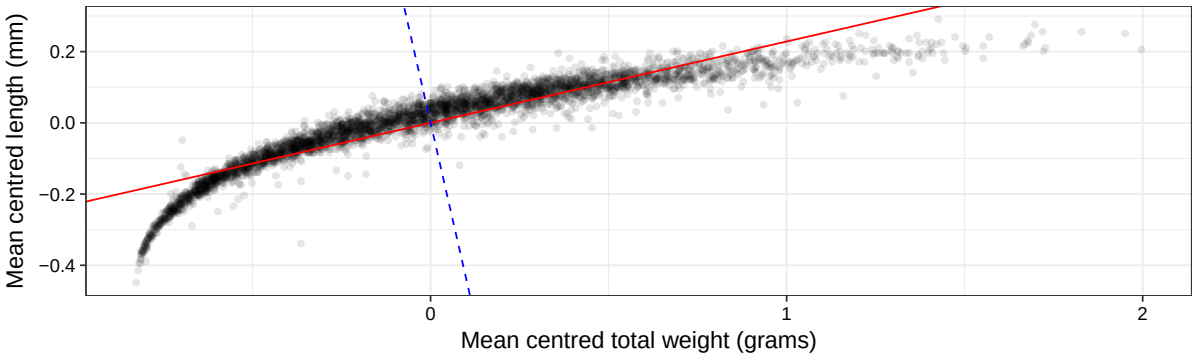
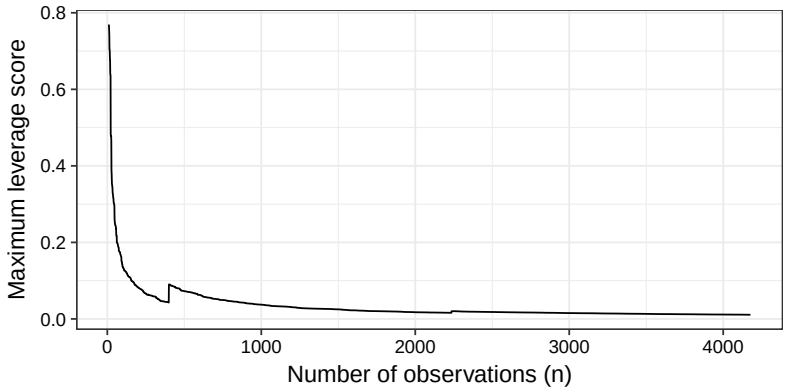


Figure 4.2: Abalone dataset ($n_{full} = 4,167$) with principal component axes.



equal to the rank of the source dataset. As we have assumed that each dataset in the sequence is of rank d, we have that $\sum_i^n \|\boldsymbol{u}_{(n)i}\|_2^2 = d$ for all n. As n grows we need the maximum contribution from a single term in the sum to tend to zero. The limiting leverage scores must satisfy an asymptotic negligibility condition, so that each individual observation provides a vanishingly small contribution to the total sum of the leverage scores.

Given a source dataset with centred columns, the leverage scores have a particularly intuitive interpretation in terms of the principal components decomposition of the source dataset. The row vector $\boldsymbol{u}_{(n)i}^{\mathsf{T}}\boldsymbol{D}_{(n)}$ gives the coordinates of observation i on the principal component axes. The elements of the vector $\boldsymbol{u}_{(n)i}$ give the coordinates of observation i in a scaled system where the variance along each principal coordinate axis is set to be one. If we think of the source dataset as a point cloud in Euclidean space, Assumption 1 essentially implies that there are no extreme outliers as n tends to infinity. Each observation must have a negligible contribution to the total variance along each principal component axis. We consider a real dataset to illustrate this concept. Figure 4.2 shows a dataset consisting of physical measurements on abalone. Each variable has been mean centred. The first and second principal components are shows as a solid and dashed lines respectively. We define a sequence of datasets $\boldsymbol{A}_{(n)}$ by taking the first n rows in the full dataset. Panel (b) shows the maximum leverage score against n, where we move through the dataset sequentially. We compute the maximum leverage score at each value of n for $n = 10, 11, \ldots, n_{full}$. The maximum leverage score appears to be heading towards zero as n increases.

Assumption 1 rules out gross outliers in the source dataset. The proof of Theorem 4.3 relies heavily on the fact that the random variables that are involved in the construction of the Clarkson-Woodruff sketch and the Hadamard sketch are bounded. The random projections use random variables that take values in $\{-1, 0, +1\}$. A second important factor is that the leverage scores of an arbitrary data matrix $\boldsymbol{A}_{(n)} = \boldsymbol{U}_{(n)}\boldsymbol{D}_{(n)}\boldsymbol{V}_{(n)}^{\mathsf{T}}$ are also bounded. For all $n \in \mathbb{N}$, $\|\boldsymbol{u}_{(n)i}\|_2^2 \leq 1$ for all $i = 1, \ldots, n$. We can thus study a sequence of bounded random matrices. Another key factor in the proof is that Hadamard matrices have a number of symmetry properties that lead to a pairwise independence structure in the random projection. These issues are discussed in more detail in Chapter 5.

4.4.3 Sketching estimators

The central limit theorem for the sketched data suggests that the results about $\boldsymbol{\beta}_S$ and $\boldsymbol{\beta}_P$ for the Gaussian sketch will also approximately hold for the Hadamard and Clarkson-Woodruff sketches for large n. In order to establish convergence of the estimators it helps to adopt an extra assumption on the sequence of source datasets.

Assumption 2:

$$\lim_{n\to\infty} n^{-1}\begin{bmatrix}\boldsymbol{y}_{(n)}^{\mathsf{T}}\boldsymbol{y}_{(n)} & \boldsymbol{y}_{(n)}^{\mathsf{T}}\boldsymbol{X}_{(n)}\\ \boldsymbol{X}_{(n)}^{\mathsf{T}}\boldsymbol{y}_{(n)} & \boldsymbol{X}_{(n)}^{\mathsf{T}}\boldsymbol{X}_{(n)}\end{bmatrix} = \boldsymbol{Q} \qquad \text{for some positive-definite matrix } \boldsymbol{Q}.$$

It is worth discussing the significance of the limiting matrix $\boldsymbol{Q}$. A useful comparison can be made to asymptotic theory for regression models, where a common assumption is that the design matrix satisfies the limit condition $n^{-1}\boldsymbol{X}_{(n)}^{\mathsf{T}}\boldsymbol{X}_{(n)} \to \boldsymbol{B}$, where $\boldsymbol{B}$ is some positive definite matrix (White, 1984; Greene, 1997). The development of asymptotic results is often eased by treating the covariates as a random sample, although this requires positing a realistic probability model for the covariates, which may be difficult. Treating the covariates as an arbitrary fixed sequence relaxes this assumption and covers more general scenarios. Although it is possible to establish asymptotic results when $n^{-1}\boldsymbol{X}_{(n)}^{\mathsf{T}}\boldsymbol{X}_{(n)}$ is not required to converge to any fixed matrix, proofs can become very technical (Fahrmeir and Tutz, 1994, Appendix A.2). Imposing a limiting value for $n^{-1}\boldsymbol{X}_{(n)}^{\mathsf{T}}\boldsymbol{X}_{(n)}$ simplifies arguments and can be seen as a compromise between making strong and weak assumptions about the covariates (Fahrmeir and Tutz, 1994, p.46). There is an analogous motivation for Assumption 2, the limiting matrix $\boldsymbol{Q}$ is present to avoid specifying a probability model for the source dataset, without overcomplicating the mathematical analysis.

Setting up a limit theorem requires a little extra care with notation. As we have a sequence of datasets $\boldsymbol{A}_{(n)}$, there is a corresponding sequence of optimal least squares coefficients $\boldsymbol{\beta}_F^{(n)}$. Similarly, there is a sequence of squared residual errors $RSS_F^{(n)}$ and model sum of squares $MSS_F^{(n)}$. As the sequence of datasets are fixed, $\boldsymbol{\beta}_F^{(n)}$, $RSS_F^{(n)}$ and $MSS_F^{(n)}$ are a deterministic sequence.

Under the assumptions 1 and 2, it is possible to establish an asymptotic result for $\boldsymbol{\beta}_S$ and $\boldsymbol{\beta}_P$.

Theorem 4.5. *Suppose that Assumptions 1 and 2 hold, $k \geq p$, and $\boldsymbol{\beta}_S$ is computed using a Hadamard or Clarkson-Woodruff sketch. Let $(\widetilde{\boldsymbol{X}}^{\mathsf{T}}\widetilde{\boldsymbol{X}})^{+}$ denote the Moore-Penrose pseudo-inverse of $(\widetilde{\boldsymbol{X}}^{\mathsf{T}}\widetilde{\boldsymbol{X}})$. Let*

$$\widetilde{\boldsymbol{H}}_{(n)} = \frac{RSS_F^{(n)}}{k}\left(\widetilde{\boldsymbol{X}}^{\mathsf{T}}\widetilde{\boldsymbol{X}}\right)^{+} \text{ and } \boldsymbol{H}_{(n)} = \frac{RSS_F^{(n)}}{k-p+1}\left(\boldsymbol{X}_{(n)}^{\mathsf{T}}\boldsymbol{X}_{(n)}\right)^{-1}.$$

Then as $n \to \infty$, convergence in distribution holds for

$$(i)[\boldsymbol{H}_{(n)}^{-1/2}(\boldsymbol{\beta}_S - \boldsymbol{\beta}_F^{(n)})|\boldsymbol{A}_{(n)}] \to \text{Student}\,(\boldsymbol{0}, \boldsymbol{I}_p,\ k-p+1)\,,$$
$$(ii)[\widetilde{\boldsymbol{H}}_{(n)}^{-1/2}(\boldsymbol{\beta}_S - \boldsymbol{\beta}_F^{(n)})\ |\boldsymbol{A}_{(n)}] \to N\,(\boldsymbol{0}, \boldsymbol{I}_p)\,.$$

The proof is in Chapter 5. For large n, we expect $\boldsymbol{\beta}_S$ to be approximately distributed as per Theorem 4.5 for both the Hadamard and Clarkson-Woodruff sketches.

It is harder to establish a comparable limit theorem for $\boldsymbol{\beta}_P^*$, due to the non-standard distribution of $\boldsymbol{\beta}_P^*$ when using a Gaussian sketch. There is no typical normalised distribution to target. Instead, we wish to show asymptotic equivalence in moments. The partially sketched estimator under the Hadamard and Clarkson-Woodruff sketches should have similar mean and variance properties to the Gaussian partially sketched estimator. An extra assumption has to be made to show convergence in moments. A sufficient condition is a stability condition on the singular values of the sketched data matrix.

Assumption 3. Let $\boldsymbol{G}$ be the Gram matrix of the scaled sketched dataset, $\boldsymbol{G} = n^{-1}\widetilde{\boldsymbol{X}}^{\mathsf{T}}\widetilde{\boldsymbol{X}}$. Assume that the sequence of source datasets is such that $E_S\left(\frac{1}{\sigma_{\min}^2(\boldsymbol{G})}\right)^2$ is finite for large enough n. This additional regularity condition enables a formal limit theorem regarding the moments of $\boldsymbol{\beta}_P^*$.

Theorem 4.6. *Suppose that Assumptions 1, 2 and 3 hold, $k > p + 3$, and $\boldsymbol{\beta}_P^*$ is computed using a Hadamard or Clarkson-Woodruff sketch. Let*

$$\boldsymbol{H}_{(n)} = \frac{(k-p-1)}{(k-p)(k-p-3)} \left(MSS_F^{(n)} (\boldsymbol{X}_{(n)}^\mathsf{T} \boldsymbol{X}_{(n)})^{-1} + \frac{(k-p+1)}{(k-p-1)} \boldsymbol{\beta}_F^{(n)} \boldsymbol{\beta}_F^{(n)\mathsf{T}} \right).$$

Then as $n \to \infty$,

$$(i)\ E_S[\boldsymbol{\beta}_P^* - \boldsymbol{\beta}_F^{(n)} | \boldsymbol{A}_{(n)}] \to \mathbf{0}.$$
$$(ii)\ \mathrm{var}_S \left(\boldsymbol{H}_{(n)}^{-1/2} (\boldsymbol{\beta}_P^* - \boldsymbol{\beta}_F^{(n)}) \ | \boldsymbol{A}_{(n)} \right) \to \boldsymbol{I}_d$$

The proof is in Chapter 5. Once again, the heavy notation may obscure the essence of the result. The subscript S is used to emphasise that the only source of randomness is the sketching matrix, and that the source dataset is fixed. The theorem suggests that the bias and variance of $\boldsymbol{\beta}_P^*$ under the Clarkson-Woodruff and Hadamard sketches should be approximately equal to that under the Gaussian sketch. Specifically, we expect equations (4.6), (4.7), and (4.8) to be good approximations for the variance of the sketched estimators using the Hadamard or Clarkson-Woodruff sketches.

The results here are meant to be useful heuristics to assess the uncertainty attached to the output of the randomised approximation algorithm. There is a need to communicate and quantify the approximation error of sketching algorithms to end users, and the asymptotic results developed in this section can be of use.

4.5 Data application

4.5.1 Human leukocyte antigen dataset

We compared the performance of the sketching estimators on a real genetic dataset taken from the UK Biobank database. We use a small extract from the data in Astle et al. (2016). The selected response variable was mean red cell volume (MCV), taken from the full blood count assay and adjusted for various technical and environmental covariates. Genome-wide imputed genotype data in expected allele dose format were available on n = 132,353 study subjects (Howie et al., 2009). We consider 1000 genetic variants in the Human leukocyte antigen (HLA) region of chromosome 6, selected so that no pair of variants had Pearson correlation of allelic scores greaterless than 0.8. The region was chosen as many associations were discovered in a genome-wide scan using univariable models; these associations were with variants with different allele frequencies, suggesting multiple distinct causal variants in the region. The aim is to perform a multivariable regression analysis to obtain variant effect size estimates that are conditional on the other variants in the region.

An early theoretical finding was that the partial sketched estimator $\boldsymbol{\beta}_P$ was biased. One thousand sketches were taken to estimate the bias $E_S(\boldsymbol{\beta}_P - \boldsymbol{\beta}_F)$ with k = 1,500. We also computed the bias corrected estimator $\boldsymbol{\beta}_P^*$ in each replication. Figure 4.3 plots the average value of the estimators against the true value of the least squares coefficient using the full dataset. The top row (a)-(c) shows results for $\boldsymbol{\beta}_P$, and the bottom row (d)-(f) shows results for $\boldsymbol{\beta}_P^*$. The first, second and third columns display the results for the Gaussian, Hadamard and Clarkson-Woodruff sketches respectively. The solid line in each panel is the identity line. The dashed line in the top row shows the theoretical bias, having slope $k/(k-p-1)$.

The results in the top row show that $\boldsymbol{\beta}_P$ is biased for each of the random projections. The bias closely matches the theoretical factor. The bottom row shows that the adjusted estimator $\boldsymbol{\beta}_P^*$ appears to be unbiased, with the mean values falling closely along the identity line.

We also compared the complete and partially sketched estimators on mean square error and the coverage of confidence intervals at k = 1,500 and k = 100,00. We did not consider a combined estimator as the small R_F^2 value would give an optimal complete sketching weight of close to zero. Table 4.2 reports

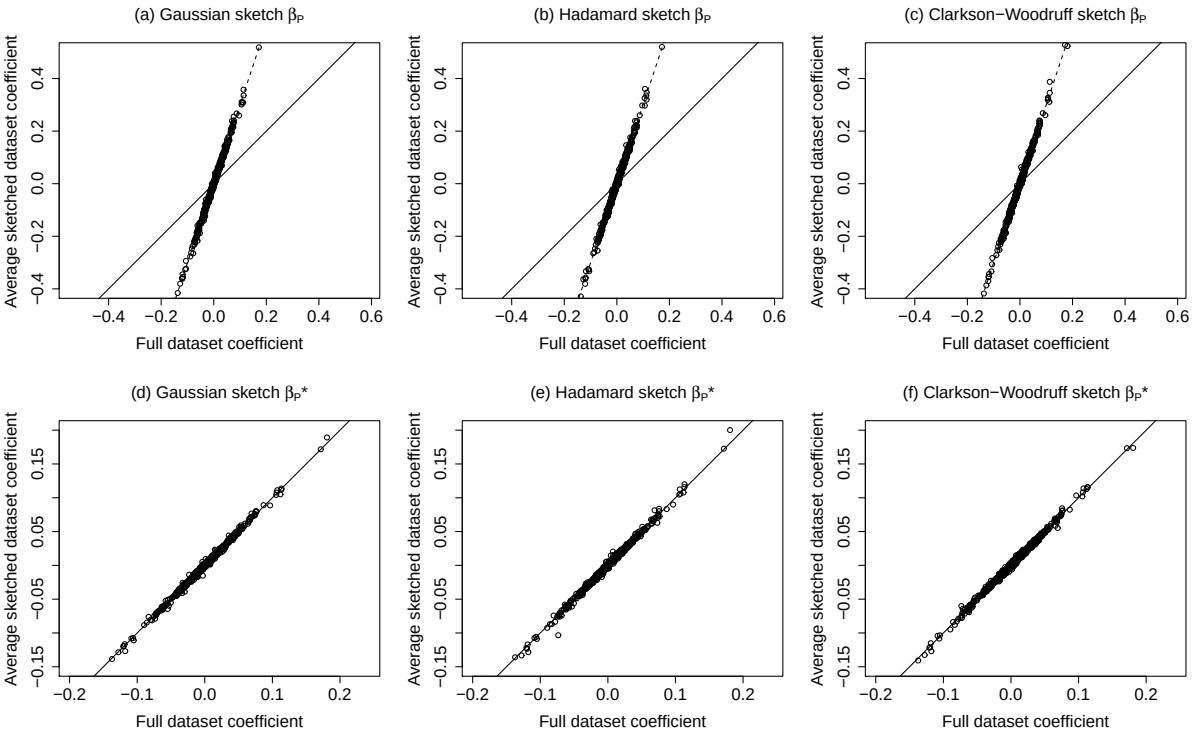


Figure 4.3: Bias of sketching estimators on the HLA dataset. Mean estimates are plotted against true values. In this scenario $n = 132353, p = 1000, k = 1500$. Solid line is the identity line and dashed line represents the theoretical bias factor.

	$k = 1,500$			$k = 10,000$		
	$\boldsymbol{\beta}_S$	$\boldsymbol{\beta}_P$	$\boldsymbol{\beta}_P^*$	$\boldsymbol{\beta}_S$	$\boldsymbol{\beta}_P$	$\boldsymbol{\beta}_P^*$
Gaussian	235 (2)	39 (0.4)	3.9 (0.04)	13.1 (0.1)	0.28 (0.002)	0.214 (0.001)
Hadamard	233 (2)	39 (0.4)	3.8 (0.04)	12.5 (0.1)	0.27 (0.002)	0.204 (0.002)
Clarkson-Woodruff	237 (2)	39 (0.5)	4.0 (0.05)	13.1 (0.1)	0.27 (0.002)	0.212 (0.002)

Table 4.2: Mean square error of sketched estimators on HLA dataset. Standard errors are in brackets.

the mean square error for each of the estimators. The signal to noise ratio is quite low for this dataset with $R_F^2 = 0.02$. The relative efficiency bound dictates that partial sketching will be much more efficient than complete sketching on this dataset. The simulation results support this idea, with $\boldsymbol{\beta}_P^*$ having a mean square error roughly sixty times smaller than $\boldsymbol{\beta}_S$ at both values of k. Results are very similar for each of the random projections, suggesting that the asymptotic approximations are reasonable for this dataset. For k = 1,500, the mean square error of $\boldsymbol{\beta}_P$ is approximately ten times that of $\boldsymbol{\beta}_P^*$. For k = 10,000, there is less of a difference, as the ratio $k/(k-p-1)$ is closer to one. The bias adjusted estimator $\boldsymbol{\beta}_P^*$ has significant advantages over $\boldsymbol{\beta}_P$ when $k/(k-p-1)$ is larger than one. Table 4.3 summarises the coverage of 95% confidence intervals for the sketched estimators. We report the overall proportion of intervals that contained the true value of the least squares estimate $\boldsymbol{\beta}_F$ over the two hundred and fifty sketches and p = 1,000 coefficients. The observed coverage is close the nominal level of 0.95 at both levels of k. The different random projections give very similar results, suggesting that the use of asymptotic approximations is again reasonable on this dataset. The intervals for the Hadamard sketch appear to be slightly conservative at k = 10,000. This may be due to the specialised random number generator used in the implementation of the Hadamard sketch. The Radamacher random variables are only four-wise independent as opposed to being mutually independent (Geppert et al., 2017, p.85).

	$k = 1,500$		$k = 10,000$	
	$\boldsymbol{\beta}_S$	$\boldsymbol{\beta}_P^*$	$\boldsymbol{\beta}_S$	$\boldsymbol{\beta}_P^*$
Gaussian	0.950	0.951	0.950	0.950
Hadamard	0.949	0.952	0.953	0.953
Clarkson-Woodruff	0.950	0.951	0.950	0.950

Table 4.3: Coverage of confidence intervals on the HLA dataset. The largest standard error is 0.001

	$\boldsymbol{\beta}_S$	$\boldsymbol{\beta}_P$	$\boldsymbol{\beta}_P^*$
Gaussian	62 (1)	15,200 (300)	14,400 (300)
Hadamard	59 (1)	14,600 (300)	14,800 (300)
Clarkson-Woodruff	59 (1)	14,700 (300)	14,100 (300)

Table 4.4: Mean square error of sketched estimators on flights dataset with $k = 5000$. Standard errors are in brackets.

	$\boldsymbol{\beta}_S$	$\boldsymbol{\beta}_P^*$
Gaussian	0.950	0.954
Hadamard	0.950	0.949
Clarkson-Woodruff	0.948	0.952

Table 4.5: Coverage of 95% confidence intervals on the flights dataset with $k = 5000$. The largest standard error is 0.004

4.5.2 Flights dataset

The sketching algorithms were also evaluated on the New York flights dataset available in the R package `nycflights13` (Wickham, 2014). Arrival delay was taken as the response, and departure delay, distance, departure time, origin and month and day were chosen to be the covariates. Rows of the dataset with missing data were omitted, leaving $n = 327{,}346$ and $p = 47$. We fit a saturated linear model with the 47 predictors. The goal was to compare the accuracy of the various sketches on real data rather than to build a statistical model for the flights dataset. We compared the mean square error of the estimators and the coverage of confidence intervals for $k = 5000$. In contrast to the HLA dataset, the flights dataset has a very high R_F^2 value of 0.99. We took one thousand sketches to compare complete and partial sketching.

Table 4.4 reports the mean square error of $\boldsymbol{\beta}_S, \boldsymbol{\beta}_P$ and $\boldsymbol{\beta}_P^*$. As expected, complete sketching has a much smaller mean square error than partial sketching. Table 4.5 summarises the coverage rates of the 95% confidence intervals. We report the overall proportion of intervals that contained the true value of the least squares estimate over the thousand sketches and $p = 47$ coefficients.

We also generated a synthetic flights dataset with an R_F^2 of close to 0.5. This was achieved by generating a synthetic response vector $\boldsymbol{y}'$ using the fitted model. The simulated response was computed as $\boldsymbol{y}' = \boldsymbol{X}\boldsymbol{\beta}_F + 15\boldsymbol{e}$, where $\boldsymbol{e}$ was the residual vector from the least squares fit $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_F$. We took one thousand sketches and computed $\boldsymbol{\beta}_S$, $\boldsymbol{\beta}_P^*$ and a weighted estimator using the optimal weight α_{opt} in each iteration. Table 4.6 reports the mean square error for each estimator. From the theoretical analysis the mean square error of the weighted estimator is expected to be roughly half that of $\boldsymbol{\beta}_S$ or $\boldsymbol{\beta}_P^*$. The results support this for each of the three different sketches.

We also assessed the finite sample behaviour of the normal approximation in Theorem 4.3 at different levels of k and p. We dropped some predictors from the full flights dataset to give smaller datasets with $p = 10$ and $p = 25$ covariates. We then took subsamples of different sizes from each of the datasets. A single subsample was taken at each value of n, so the same subsampled dataset was being sketched each

	$\boldsymbol{\beta}_S$	$\boldsymbol{\beta}_P^*$	$\boldsymbol{\beta}_C$ with $\alpha = \alpha_{\text{opt}}$
Gaussian	13,500 (300)	13,700 (300)	7,000 (150)
Hadamard	13,100 (300)	14,300 (300)	6,900 (150)
Clarkson-Woodruff	13,600 (300)	14,600 (300)	7,000 (150)

Table 4.6: Mean square error of sketched estimators on synthetic flights dataset with $k = 5000$. Standard errors are in brackets.

time. One thousand sketches were taken of each dataset at different values of k. We tested the joint multivariate normality of $[\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}]$ and the normality of the sketched residual $\widetilde{\boldsymbol{e}} = \boldsymbol{S}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_F)$. The squared Mahalanobis distance of the sketched observations was compared to the theoretical χ^2-distribution. As n increases the rejection rate is expected to fall to the type one error rate of 0.05. Figure 4.4 plots the proportion of times the null hypothesis of normality is rejected against the size of the source dataset.

The Hadamard sketch appears to have a much faster rate of convergence than the Clarkson-Woodruff sketch. When using a Hadamard sketch, each row in the sketched dataset is a linear combination of n observations from the source dataset. When using a Clarkson-Woodruff sketch, each row in the sketched dataset is expected to be a combination of only n/k observations from the source dataset. As such, n/k must be large for the normal approximation to hold. As expected, the rejection rate for the Clarkson-Woodruff sketch increases with k, but remains stable for the Hadamard sketch. In Figure 4.4 the rejection rate for the Clarkson-Woodruff sketch increases with p. The Hadamard sketch seems to be less sensitive to the number of covariates. The extra $\log k$ computation cost associated with the Hadamard sketch (Table 4.1) appears to have the benefit of accelerated convergence to normality. Even though joint normality may not be holding for the Clarkson-Woodruff sketch for the flights dataset, the coverage of the confidence intervals is still very good. As $\widetilde{\boldsymbol{y}} = \widetilde{\boldsymbol{X}}\boldsymbol{\beta}_F + \widetilde{\boldsymbol{e}}$, normality of the sketched residual is perhaps sufficient in justifying the approximate confidence intervals given by Theorem 4 (ii). The sketched residual converges much more quickly than the full sketched data matrix, which perhaps explains the good coverage properties of the confidence intervals for $\boldsymbol{\beta}_S$ in Table 4.5.

4.6 Discussion

Sketching algorithms have emerged in the computer science community as a powerful device for the analysis of massive datasets (Mahoney and Drineas, 2016). Sketched regression algorithms use random projections to reduce the size of the original dataset, the sketched dataset is then used to estimate the optimal least squares coefficients. Most existing theory for sketched regression is from an algorithmic worst case perspective and connects with random matrix theory and computational geometry (Raskutti and Mahoney, 2014; Thanei et al., 2017). In this chapter we have provided a complementary statistical perspective and derived new tools for assessing the uncertainty attached to sketched estimators as well as guidelines for choosing between competing sketching algorithms.

Iterative methods, in particular stochastic gradient descent have not been mentioned so far. For large n regression problems, stochastic gradient descent will produce iterates that converge to $\boldsymbol{\beta}_F$ under very mild conditions. Comparisons between single pass sketching and stochastic gradient methods are difficult, as the two techniques are not formulated for the exact same purpose. Single pass sketching algorithms are designed to return an approximate solution in finite time with probabilistically controlled error, whereas stochastic gradient methods are designed to converge to the exact solution asymptotically. It is perhaps more appropriate to compare stochastic gradient descent to iterative sketching methods, as iterative sketching algorithms also come with convergence guarantees to $\boldsymbol{\beta}_F$ (Pilanci and Wainwright, 2016; Gower and Richtrik, 2015). Iterative sketching methods make use of approximate second order information that can lead to a potential improvement compared to first order stochastic gradient methods (Roosta-Khorasani and Mahoney, 2016). Our focus has been on characterising the approximation error

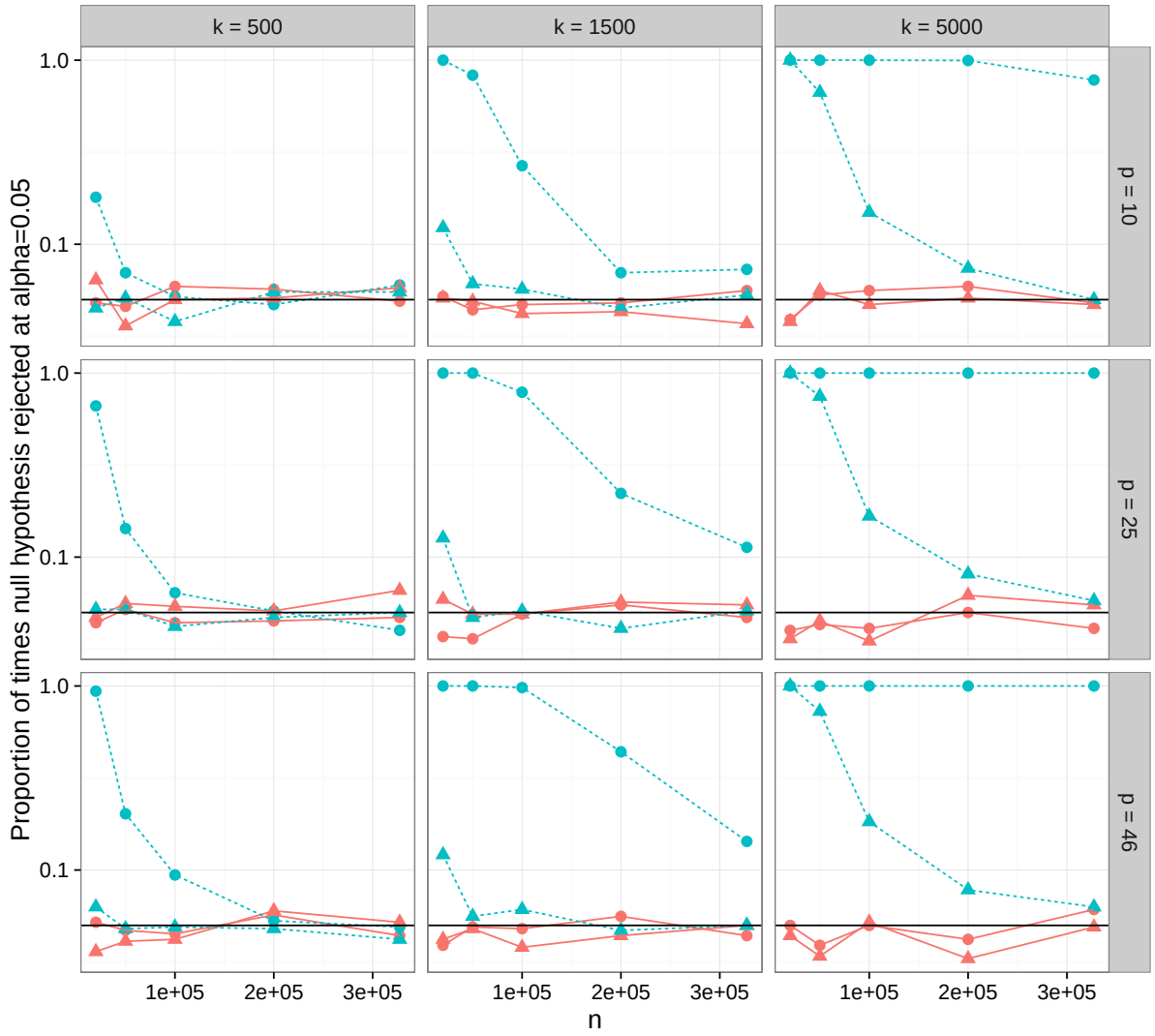


Figure 4.4: Proportion of times null hypothesis of normality is rejected against size of the source dataset (n) for the Hadamard (solid line) and Clarkson-Woodruff sketches (dashed line). Results for tests of the sketched residual vector $\widetilde{\boldsymbol{e}} = \boldsymbol{S}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_F)$ are plotted as triangles ($\triangle$), and results for tests of the entire sketched dataset $[\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}]$ are plotted as circles ($\circ$). The horizontal line gives the type 1 error of 0.05. The y-axis is on a log scale.

attached to single pass sketching estimators. The spectral theory developed in Chapter 6 has applications to iterative sketching algorithms, although that is not the focus of the chapter.

There has been recent work in adapting sketching methods for statistical inference in large datasets, building from the worst case bounds in the computer science literature. Geppert et al. (2017) and Bardenet and Maillard (2015) investigate sketching algorithms for Bayesian regression, and derive bounds on the difference between the sketched posterior distribution and the full data posterior distribution. Yang et al. (2015b) consider sketched penalised regression, and give bounds between the sketched solution and the full data solution similar to the results in section 4.2.1. Only complete sketching is considered in the aforementioned work. The results on the advantages of partial sketching in this paper could motivate adaptations that make use of the exact marginal associations $\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$.

Sketching ideas have been used to develop methods for approximate non-linear regression (Avron et al., 2014; Banerjee et al., 2013). A related branch of work uses random projections to reduce the number of predictors in regression and classification problems (Shah and Meinshausen, 2013; Cannings and Samworth, 2015; Guhaniyogi and Dunson, 2015). Sketching can also be used to ensure privacy when sharing sensitive datasets (Zhou et al., 2009).

Acknowledgement

This work has been conducted using the UK Biobank resource under applications number 13745. Many thanks to Rajen Shah for helpful discussions.

5

Proofs regarding sketching algorithms

Summary

This Chapter contains proofs for the results in Chapter 4. We restate the major results before giving proofs.

5.1 Proof of Theorem 4.1 (Worst case bound for partial sketching)

Theorem. *Suppose that $\widetilde{\boldsymbol{X}}$ is an ϵ-subspace embedding of $\boldsymbol{X}$ with $0 < \epsilon < 0.5$. Then the following bound holds,*

$$\|\boldsymbol{\beta}_P - \boldsymbol{\beta}_F\|_2^2 \leq \frac{4\epsilon^2}{\sigma_{\min}^2(\boldsymbol{X})} MSS_F.$$

Let the singular value decomposition of $\boldsymbol{X}$ be given by $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\mathsf{T}$. The singular value decomposition will help to simplify expressions in later working. If the sketching matrix $\boldsymbol{S}$ is an ϵ-subspace embedding for the source dataset with $0 < \epsilon < 1$, then $\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}$ is necessarily invertible. The expression for $\boldsymbol{\beta}_P$ can then be simplified to

$$\begin{aligned}
\boldsymbol{\beta}_P &= \boldsymbol{V}\boldsymbol{D}^{-1}(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})^{-1}\boldsymbol{D}^{-1}\boldsymbol{V}^\mathsf{T}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} \\
&= \boldsymbol{V}\boldsymbol{D}^{-1}(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})^{-1}\boldsymbol{D}^{-1}\boldsymbol{V}^\mathsf{T}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\mathsf{T}\boldsymbol{y} \\
&= \boldsymbol{V}\boldsymbol{D}^{-1}(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})^{-1}\boldsymbol{U}^\mathsf{T}\boldsymbol{y}.
\end{aligned}$$

Similarly, $\boldsymbol{\beta}_F$ can be written as $\boldsymbol{\beta}_F = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^\mathsf{T}\boldsymbol{y}$. The Euclidean norm of the approximation error can thus be expressed as

$$\begin{aligned}
\|\boldsymbol{\beta}_P - \boldsymbol{\beta}_F\|_2 &= \|\boldsymbol{V}\boldsymbol{D}^{-1}(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})^{-1}\boldsymbol{U}^\mathsf{T}\boldsymbol{y} - \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^\mathsf{T}\boldsymbol{y}\|_2 \\
&= \|\left\{\boldsymbol{V}\boldsymbol{D}^{-1}(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})^{-1} - \boldsymbol{V}\boldsymbol{D}^{-1}\right\}\boldsymbol{U}^\mathsf{T}\boldsymbol{y}\|_2 \\
&= \|\left\{\boldsymbol{V}\boldsymbol{D}^{-1}[(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})^{-1} - \boldsymbol{I}_p]\right\}\boldsymbol{U}^\mathsf{T}\boldsymbol{y}\|_2.
\end{aligned}$$

The model sum of squares can be written as

$$\begin{aligned}
MSS_F &= \|\boldsymbol{X}\boldsymbol{\beta}_F\|_2^2 \\
&= \|\boldsymbol{X}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^\mathsf{T}\boldsymbol{y}\|_2^2 \\
&= \|\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\mathsf{T}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^\mathsf{T}\boldsymbol{y}\|_2^2 \\
&= \|\boldsymbol{U}\boldsymbol{U}^\mathsf{T}\boldsymbol{y}\|_2^2 \\
&= \|\boldsymbol{U}^\mathsf{T}\boldsymbol{y}\|_2^2. \qquad (5.1)
\end{aligned}$$

The final line uses the fact that $\boldsymbol{U}^\mathsf{T}\boldsymbol{U} = \boldsymbol{I}_p$. Using the matrix norm induced by the Euclidean norm and the usual Euclidean norm for vectors we can form an upper bound on the error.

$$\begin{aligned}\|\boldsymbol{\beta}_P - \boldsymbol{\beta}_F\|_2 &\leq \|\boldsymbol{V}\boldsymbol{D}^{-1}\left\{(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})^{-1} - \boldsymbol{I}_p\right\}\|_2\|\boldsymbol{U}^\mathsf{T}\boldsymbol{y}\|_2 \\ &\leq \|\boldsymbol{V}\boldsymbol{D}^{-1}\|_2\|\boldsymbol{U}^\mathsf{T}\boldsymbol{y}\|_2\|(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})^{-1} - \boldsymbol{I}_p\|_2 \\ &= \frac{MSS_F^{1/2}}{\sigma_{\min}(\boldsymbol{X})}\|(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})^{-1} - \boldsymbol{I}_p\|_2. \end{aligned} \tag{5.2}$$

It remains to upper bound the maximum singular value of the matrix $(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})^{-1} - \boldsymbol{I}_p$. Let $\boldsymbol{M} = \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}$. The maximum absolute value of the singular values of $(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})^{-1} - \boldsymbol{I}_p$ will be given by $\max(|1/\sigma_{\min}(\boldsymbol{M}) - 1|, |1/\sigma_{\max}(\boldsymbol{M}) - 1|)$, where $\sigma_{\min}(\boldsymbol{M})$ is the minimum singular value of $\boldsymbol{M}$, and $\sigma_{\max}(\boldsymbol{M})$ is the maximum singular value of $\boldsymbol{M}$. If $\boldsymbol{S}$ is an ϵ-subspace embedding for the source covariate matrix $\boldsymbol{X}$ then it must hold that $\sigma_{\min}(\boldsymbol{M}) \geq 1 - \epsilon$, and $\sigma_{\max}(\boldsymbol{M}) \leq 1 + \epsilon$ (Woodruff, 2014, p.11). As such, $\max(|1/\sigma_{\min}(\boldsymbol{M}) - 1|, |1/\sigma_{\max}(\boldsymbol{M}) - 1|) \leq |1/(1-\epsilon) - 1|$. It is simple to show that over the interval $0 \leq \epsilon \leq 0.5$, $|1/(1-\epsilon) - 1| \leq 2\epsilon$. This results in an upper bound on the singular value of interest,

$$\begin{aligned}\|(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})^{-1} - \boldsymbol{I}_p\|_2 &\leq |1/(1-\epsilon) - 1| \\ &\leq 2\epsilon.\end{aligned}$$

Substituting this back into (5.2) gives that under the condition that $\epsilon < 0.5$

$$\|\boldsymbol{\beta}_P - \boldsymbol{\beta}_F\|_2 \leq \frac{MSS_F^{1/2}}{\sigma_{\min}(\boldsymbol{X})} \times 2\epsilon.$$

Squaring both sides gives the final result, that if $\epsilon < 0.5$

$$\|\boldsymbol{\beta}_P - \boldsymbol{\beta}_F\|_2^2 \leq \frac{4\epsilon^2}{\sigma_{\min}^2(\boldsymbol{X})}MSS_F.$$

5.2 Proof of Theorem 4.2 (Hierarchical model for the Gaussian sketch)

Theorem. *Suppose $\boldsymbol{\beta}_S$ is computed using a Gaussian sketch and $k > p+1$. The conditional distribution of $\boldsymbol{\beta}_S$ is*

$$(i)\ \boldsymbol{\beta}_S|\widetilde{\boldsymbol{X}}, \boldsymbol{y}, \boldsymbol{X} \sim N\left(\boldsymbol{\beta}_F, \frac{n\sigma_F^2}{k}\left(\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}}\right)^{-1}\right).$$

The marginal distribution of $\boldsymbol{\beta}_S$ is

$$(ii)\ \boldsymbol{\beta}_S|\boldsymbol{y}, \boldsymbol{X} \sim \text{Student}\left(\boldsymbol{\beta}_F, \frac{n\sigma_F^2}{k-p+1}\left(\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\right)^{-1},\ k-p+1\right).$$

We use the following lemma about the Normal Inverse-Wishart distribution in many of our results (Gelman et al., 2014, p.73). Suppose that $\boldsymbol{\Sigma}$ is a random $d \times d$ matrix and $\boldsymbol{y}$ is a d-dimensional random vector from the following hierarchical model

$$\begin{aligned}\boldsymbol{y}|\Sigma &\sim N\left(\boldsymbol{\mu}, \Sigma/\kappa\right), \\ \Sigma &\sim \text{Inv-Wishart}(\Lambda, \nu),\end{aligned}$$

where Λ is a $d \times d$ scale matrix, ν is a scalar giving degrees of freedom, and κ is a scaling constant. Then marginally,

$$\boldsymbol{y} \sim \text{Student}(\boldsymbol{\mu}, \Lambda/(\kappa(\nu - d + 1)), \nu - d + 1).$$

Theorem 4.2 follows from setting $\boldsymbol{\mu} = \boldsymbol{\beta}_F$, $\Sigma = (\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})^{-1}$, $\kappa = k/RSS_F$, $\Lambda = k(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}$, $\nu = k$ and $d = p$.

5.3 Variance for partial sketching

Using a Gaussian sketch of size k where $k > p + 3$, the standard partial sketching estimator $\boldsymbol{\beta}_P$ has variance

$$\text{var}(\boldsymbol{\beta}_P) = \frac{k^2}{(k-p)(k-p-1)(k-p-3)} \left(MSS_F(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1} + \frac{(k-p+1)}{(k-p-1)}\boldsymbol{\beta}_F\boldsymbol{\beta}_F^\mathsf{T} \right). \tag{5.3}$$

The bias corrected partial sketching estimator $\boldsymbol{\beta}_P^*$ has variance

$$\text{var}(\boldsymbol{\beta}_P^*) = \frac{(k-p-1)}{(k-p)(k-p-3)} \left(MSS_F(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1} + \frac{(k-p+1)}{(k-p-1)}\boldsymbol{\beta}_F\boldsymbol{\beta}_F^\mathsf{T} \right). \tag{5.4}$$

We now prove (5.3) and (5.4).

Let the singular value decomposition of $\boldsymbol{X}$ be given by $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\mathsf{T}$. The singular value decomposition will help to simplify expressions in later working. The sketched Gram matrix has the form $\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\mathsf{T}$. As $\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U} \sim \text{Wishart}(k, \boldsymbol{I}_p/k)$, the matrix $\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}$ is almost surely invertible. The inverse Gram matrix can then be written as

$$\begin{aligned}(\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})^{-1} &= [\boldsymbol{D}\boldsymbol{V}^\mathsf{T}]^{-1}(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})^{-1}[\boldsymbol{V}\boldsymbol{D}]^{-1} \\ &= \boldsymbol{V}\boldsymbol{D}^{-1}(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})^{-1}\boldsymbol{D}^{-1}\boldsymbol{V}^\mathsf{T}.\end{aligned}$$

The expression for $\boldsymbol{\beta}_P$ can then be simplified to

$$\begin{aligned}\boldsymbol{\beta}_P &= \boldsymbol{V}\boldsymbol{D}^{-1}(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})^{-1}\boldsymbol{D}^{-1}\boldsymbol{V}^\mathsf{T}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} \\ &= \boldsymbol{V}\boldsymbol{D}^{-1}(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})^{-1}\boldsymbol{D}^{-1}\boldsymbol{V}^\mathsf{T}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\mathsf{T}\boldsymbol{y} \\ &= \boldsymbol{V}\boldsymbol{D}^{-1}(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})^{-1}\boldsymbol{U}^\mathsf{T}\boldsymbol{y}.\end{aligned}$$

Let $\boldsymbol{M} = (\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})^{-1}$. We know that $\boldsymbol{M} \sim \text{Inverse-Wishart}(k, k\boldsymbol{I}_p)$. Properties of the Inverse-Wishart distribution give that that for $i = 1, \ldots, p$,

$$\text{var}(M_{ii}) = \frac{2k^2}{(k-p-1)^2(k-p-3)}. \tag{5.5}$$

Additionally, for $i, j = 1, \ldots, p$, where $j \neq i$

$$\text{var}(M_{ij}) = \frac{k^2(k-p-1)}{(k-p)(k-p-1)^2(k-p-3)}. \tag{5.6}$$

Finally we have that for $i, j = 1, \ldots, p$, $i \neq j$,

$$\text{cov}(M_{ij}, M_{ji}) = \frac{k^2(k-p-1)}{(k-p)(k-p-1)^2(k-p-3)}, \tag{5.7}$$

$$\text{cov}(M_{ii}, M_{jj}) = \frac{2k^2}{(k-p)(k-p-1)^2(k-p-3)}. \tag{5.8}$$

All other covariances $\text{cov}(M_{ij}, M_{br})$ are equal to zero unless they reduce to the cases in (5.7) or (5.8). Let $\boldsymbol{z} = \boldsymbol{U}^\mathsf{T}\boldsymbol{y}$. Let $\boldsymbol{W} = \text{cov}\left(\boldsymbol{M}\boldsymbol{U}^\mathsf{T}\boldsymbol{y}\right) = \text{cov}(\boldsymbol{M}\boldsymbol{z})$. The elements of $\boldsymbol{W}$ can be determined using the properties in equations (5.5) to (5.8). Starting with the diagonal entries,

$$\begin{aligned}W_{ii} &= \text{var}\left(\sum_{j=1}^{p} M_{ij}z_j\right) \\ &= \sum_{j=1}^{p} z_j^2\text{var}(M_{ij}) + \sum_{j=1}^{p}\sum_{w\neq j}^{p} z_j z_w \text{cov}(M_{ij}, M_{iw}).\end{aligned}$$

As $\text{cov}(M_{ij}, M_{iw})$ is equal to zero for all $w \neq j$ this simplifies to

$$\begin{aligned}W_{ii} &= \text{var}\left(\sum_{j=1}^{p} M_{ij}z_j\right) \\ &= \sum_{j=1}^{p} z_j^2 \text{ var}(M_{ij}).\end{aligned}$$

It is helpful to split the sum into two pieces, a single term for $j = i$ and then a sum over the remaining indices. Grouping terms leads to an expression involving the model sum of squares MSS_F.

$$\begin{aligned}
W_{ii} &= z_i^2 \frac{2k^2}{(k-p-1)^2(k-p-3)} + \sum_{j=1, j\neq i}^{p} z_j^2 \frac{k^2(k-p-1)}{(k-p)(k-p-1)^2(k-p-3)} \\
&= z_i^2 \frac{2k^2(k-p)}{(k-p)(k-p-1)^2(k-p-3)} + \sum_{j=1, j\neq i}^{p} z_j^2 \frac{k^2(k-p-1)}{(k-p)(k-p-1)^2(k-p-3)} \\
&= z_i^2 \frac{2k^2(k-p-1)+2k^2}{(k-p)(k-p-1)^2(k-p-3)} + \sum_{j=1, j\neq i}^{p} z_j^2 \frac{k^2(k-p-1)}{(k-p)(k-p-1)^2(k-p-3)} \\
&= \frac{k^2(k-p-1)}{(k-p)(k-p-1)^2(k-p-3)} \sum_{j=1}^{p} z_j^2 + \frac{k^2(k-p-1)+2k^2}{(k-p)(k-p-1)^2(k-p-3)} z_i^2 \\
&= \frac{k^2(k-p-1)}{(k-p)(k-p-1)^2(k-p-3)} MSS_F + \frac{k^2(k-p+1)}{(k-p)(k-p-1)^2(k-p-3)} z_i^2.
\end{aligned}$$

In the second line the first term is modified to have the same denominator as the remainder sum. In the third line we add and subtract by $2k^2$ so that the numerator in the first term matches the numerator in the remainder sum. This allows the z_j terms to be grouped into a sum over the full set of indexes $j = 1, \ldots, p$ in the third line. The fourth line uses the fact that $\sum_{j=1}^{p} z_j^2 = \boldsymbol{z}^\mathsf{T}\boldsymbol{z} = MSS_F$. This was shown in the proof of Theorem 1 (5.1). For the off diagonal entries W_{ib} where $b \neq i$,

$$\begin{aligned}
W_{ib} &= \operatorname{cov}\left(\sum_{j=1}^{p} M_{ij} z_j, \sum_{r=1}^{p} M_{br} z_r \right) \\
&= \sum_{j=1}^{p} \sum_{r=1}^{p} z_j z_r \operatorname{cov}(M_{ij}, M_{br}).
\end{aligned}$$

Now $\operatorname{cov}(M_{ij}, M_{br})$ is only nonzero for $\operatorname{cov}(M_{ib}, M_{bi})$ and $\operatorname{cov}(M_{ii}, M_{bb})$. Using (5.7) and (5.8) we obtain

$$\begin{aligned}
W_{ib} &= z_i z_b \operatorname{cov}(M_{ib}, M_{bi}) + z_i z_b \operatorname{cov}(M_{ii}, M_{bb}) \\
&= \frac{k^2(k-p-1)}{(k-p)(k-p-1)^2(k-p-3)} z_i z_b + \frac{2k^2}{(k-p)(k-p-1)^2(k-p-3)} z_i z_b \\
&= \frac{k^2(k-p+1)}{(k-p)(k-p-1)^2(k-p-3)} z_i z_b.
\end{aligned}$$

The entire covariance matrix $\boldsymbol{W}$ can therefore be written compactly as

$$\begin{aligned}
\boldsymbol{W} &= \frac{k^2(k-p-1)}{(k-p)(k-p-1)^2(k-p-3)} (MSS_F \boldsymbol{I}_p) + \frac{k^2(k-p+1)}{(k-p)(k-p-1)^2(k-p-3)} \boldsymbol{z}\boldsymbol{z}^T \\
&= \frac{k^2(k-p-1)}{(k-p)(k-p-1)^2(k-p-3)} \left(MSS_F \boldsymbol{I}_p + \frac{(k-p+1)}{(k-p-1)} \boldsymbol{z}\boldsymbol{z}^T \right). \\
&= \frac{k^2}{(k-p)(k-p-1)(k-p-3)} \left(MSS_F \boldsymbol{I}_p + \frac{(k-p+1)}{(k-p-1)} \boldsymbol{z}\boldsymbol{z}^T \right).
\end{aligned}$$

Now $\boldsymbol{\beta}_P = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{M}\boldsymbol{z}$. Therefore $\operatorname{var}(\boldsymbol{\beta}_P) = \boldsymbol{V}\boldsymbol{D}^{-1}\operatorname{var}(\boldsymbol{M}\boldsymbol{z})\boldsymbol{D}^{-1}\boldsymbol{V}^\mathsf{T} = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{W}\boldsymbol{D}^{-1}\boldsymbol{V}^\mathsf{T}$. The variance of $\boldsymbol{\beta}_P$ is then a linear function of $\boldsymbol{W}$,

$$\begin{aligned}
\operatorname{var}(\boldsymbol{\beta}_P) &= \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{W}\boldsymbol{D}^{-1}\boldsymbol{V}^\mathsf{T} \\
&= \boldsymbol{V}\boldsymbol{D}^{-1} \frac{k^2}{(k-p)(k-p-1)(k-p-3)} \left(MSS_F \boldsymbol{I}_p + \frac{(k-p+1)}{(k-p-1)} \boldsymbol{z}\boldsymbol{z}^T \right) \boldsymbol{D}^{-1}\boldsymbol{V}^\mathsf{T} \\
&= \frac{k^2}{(k-p)(k-p-1)(k-p-3)} MSS_F (\boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^\mathsf{T}) + \\
&\quad \frac{k^2(k-p+1)}{(k-p)(k-p-1)^2(k-p-3)} \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{z}\boldsymbol{z}^\mathsf{T}\boldsymbol{D}^{-1}\boldsymbol{V}^\mathsf{T}
\end{aligned} \tag{5.9}$$

Recall that $\boldsymbol{z} = \boldsymbol{U}^\mathsf{T}\boldsymbol{y}$ and

$$\boldsymbol{\beta}_F = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} \tag{5.10}$$
$$= \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^\mathsf{T}\boldsymbol{y} \tag{5.11}$$
$$= \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{z}. \tag{5.12}$$

The term $\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{z}$ appears in (5.9). Substituting (5.12) into (5.9) gives

$$\text{var}(\boldsymbol{\beta}_P) = \frac{k^2}{(k-p)(k-p-1)(k-p-3)} MSS_F(\boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^\mathsf{T}) + \frac{k^2(k-p+1)}{(k-p)(k-p-1)^2(k-p-3)}\boldsymbol{\beta}_F\boldsymbol{\beta}_F^\mathsf{T}.$$

A final simplification can be made by noting that $(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1} = \boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^\mathsf{T}$ giving

$$\begin{aligned}\text{var}(\boldsymbol{\beta}_P) &= \frac{k^2}{(k-p)(k-p-1)(k-p-3)} MSS_F(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1} + \frac{k^2(k-p+1)}{(k-p)(k-p-1)^2(k-p-3)}\boldsymbol{\beta}_F\boldsymbol{\beta}_F^\mathsf{T} \\ &= \frac{k^2}{(k-p)(k-p-1)(k-p-3)}\left(MSS_F(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1} + \frac{(k-p+1)}{(k-p-1)}\boldsymbol{\beta}_F\boldsymbol{\beta}_F^\mathsf{T}\right).\end{aligned}$$

The variance of $\boldsymbol{\beta}_P^* = [(k-p-1)/k]\boldsymbol{\beta}_P$ is then

$$\begin{aligned}\text{var}(\boldsymbol{\beta}_P^*) &= \left(\frac{k-p-1}{k}\right)^2 \frac{k^2(k-p-1)}{(k-p)(k-p-1)^2(k-p-3)}\left(MSS_F(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1} + \frac{(k-p+1)}{(k-p-1)}\boldsymbol{\beta}_F\boldsymbol{\beta}_F^\mathsf{T}\right) \\ &= \frac{(k-p-1)}{(k-p)(k-p-3)}\left(MSS_F(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1} + \frac{(k-p+1)}{(k-p-1)}\boldsymbol{\beta}_F\boldsymbol{\beta}_F^\mathsf{T}\right).\end{aligned} \tag{5.13}$$

5.4 Combined estimator results

We first show that $\boldsymbol{\beta}_P^*$ and $\boldsymbol{\beta}_S$ are uncorrelated. We again avoid notation explicitly conditioning on the source dataset $[\boldsymbol{y}, \boldsymbol{X}]$ in every step, which is always treated as fixed. The covariance between $\boldsymbol{\beta}_P^*$ and $\boldsymbol{\beta}_S$ computed from the same sketch can be shown to be zero. Using the definition of covariance, and taking iterated expectations

$$\begin{aligned}\text{cov}(\boldsymbol{\beta}_P^*, \boldsymbol{\beta}_S) &= \mathbb{E}\left[(\boldsymbol{\beta}_P^* - \boldsymbol{\beta}_F)(\boldsymbol{\beta}_S - \boldsymbol{\beta}_F)^\mathsf{T}\right] \\ &= \mathbb{E}\left\{\mathbb{E}\left[(\boldsymbol{\beta}_P^* - \boldsymbol{\beta}_F)(\boldsymbol{\beta}_S - \boldsymbol{\beta}_F)^\mathsf{T} \mid \widetilde{\boldsymbol{X}}\right]\right\}.\end{aligned}$$

Recall the hierarchical model for complete sketching,

$$\widetilde{\boldsymbol{y}} \mid \widetilde{\boldsymbol{X}} \sim N\left(\widetilde{\boldsymbol{X}}\boldsymbol{\beta}_F, \frac{RSS_F}{k}\boldsymbol{I}_k\right).$$

Equivalently,

$$\widetilde{\boldsymbol{y}} \mid \widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{X}}\boldsymbol{\beta}_F + \widetilde{\boldsymbol{e}},$$

where $\widetilde{\boldsymbol{e}} \mid \widetilde{\boldsymbol{X}} \sim N(\boldsymbol{0}, \frac{RSS_F}{k}\boldsymbol{I}_k)$. So

$$\boldsymbol{\beta}_S \mid \widetilde{\boldsymbol{X}}, \boldsymbol{y}, \boldsymbol{X} = \boldsymbol{\beta}_F + (\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{e}}.$$

Substituting back into the expression for the covariance,

$$\begin{aligned}\text{cov}(\boldsymbol{\beta}_P^*, \boldsymbol{\beta}_S) &= \mathbb{E}\left\{\mathbb{E}\left[(\boldsymbol{\beta}_P^* - \boldsymbol{\beta}_F)(\boldsymbol{\beta}_F + (\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{e}} - \boldsymbol{\beta}_F)^\mathsf{T} \mid \widetilde{\boldsymbol{X}}\right]\right\} \\ &= \mathbb{E}\left\{\left[(\boldsymbol{\beta}_P^* - \boldsymbol{\beta}_F)(\boldsymbol{\beta}_F + (\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}^\mathsf{T}\mathbb{E}[\widetilde{\boldsymbol{e}} \mid \widetilde{\boldsymbol{X}}] - \boldsymbol{\beta}_F)^\mathsf{T} \mid \widetilde{\boldsymbol{X}}\right]\right\} \\ &= \mathbb{E}\left\{\left[(\boldsymbol{\beta}_P^* - \boldsymbol{\beta}_F)(\boldsymbol{\beta}_F - \boldsymbol{\beta}_F)^\mathsf{T} \mid \widetilde{\boldsymbol{X}}\right]\right\} \\ &= \mathbb{E}\left\{\left[(\boldsymbol{\beta}_P^* - \boldsymbol{\beta}_F)\boldsymbol{0}^\mathsf{T} \mid \widetilde{\boldsymbol{X}}\right]\right\} \\ &= \boldsymbol{0}_{p\times p}.\end{aligned}$$

The final line may be obtained by noting that $\mathbb{E}[\boldsymbol{\beta}_P^*] = \boldsymbol{\beta}_F$. Simple calculus shows that the value which minimises the expected mean square error $\mathbb{E}_S(\|\boldsymbol{\beta}_C - \boldsymbol{\beta}_F\|_2^2 \mid \boldsymbol{y}, \boldsymbol{X})$ is

$$\alpha_{\text{opt}} = \frac{\text{trace}(\text{var}(\boldsymbol{\beta}_P^*))}{\text{trace}(\text{var}(\boldsymbol{\beta}_P^*)) + \text{trace}(\text{var}(\boldsymbol{\beta}_S))}.$$

5.5 Proof of Theorem 4.4 (central limit theorem under asymptotic negligibility condition)

A triangular array of random variables is a useful structure for studying weak convergence. To establish a triangular array, define for every $n \in \mathbb{N}$ a collection of random variables $Z_{n1}, Z_{n2}, \ldots, Z_{nr_n}$. There are r_n random variables in row n of the array. Suppose that $r_n = n$. Visually we can represent the first three rows of the array as

$$\begin{array}{ccccc} Z_{11} & & & & \\ Z_{21} & Z_{22} & & & \\ Z_{31} & Z_{32} & Z_{33} & & \end{array}$$

Theorem (Billingsley, 1999)**.** *For each $n \in \mathbb{N}$, let $Z_{n1}, Z_{n2}, \ldots, Z_{nr_n}$ be a sequence of independent random variables with $\mathbb{E}(Z_{ni}) = 0$ and $\text{var}(Z_{ni}) = \sigma_{ni}^2$ for $i = 1, \ldots, r_n$. Let $s_n^2 = \sum_{i=1}^{r_n} \sigma_{ni}^2$ and assume that $r_n \to \infty$ as $n \to \infty$. Suppose that we can form a sequence of upper bounds $(K_n)_{n\in\mathbb{N}}$ such that*

$$|Z_{ni}| \leq K_n \text{ almost surely for } i = 1, \ldots, r_n.$$

Then if $K_n/s_n \to 0$ as $n \to \infty$ we have the convergence in distribution

$$\frac{1}{s_n} \sum_{i=1}^{r_n} Z_{ni} \xrightarrow{d} N(0, 1)$$

Lindeberg's condition is a critical component in establishing asymptotic normality. We state Lindeberg's condition for triangular arrays of random variables.

Definition 5.1 (Lindeberg's condition)**.** *For each $n \in \mathbb{N}$, let $Z_{n1}, Z_{n2}, \ldots, Z_{nr_n}$ be a sequence of random variables with $\mathbb{E}(Z_{ni}) = 0$ and $\text{var}(Z_{ni}) = \sigma_{ni}^2$ for $i = 1, \ldots, r_n$. Let $s_n^2 = \sum_{i=1}^{n} \sigma_{ni}^2$ and suppose that $r_n \to \infty$ as $n \to \infty$. The random variables are said to satisfy Lindeberg's condition if for all $\eta > 0$,*

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^{r_n} \mathbb{E}(Z_{ni}^2 \mathbb{1}_{\{|Z_{ni}|>\eta s_n\}}) = 0. \tag{5.14}$$

The triangular array of random variables does not have to have independent random variables in each row in order to satisfy the condition. The general form of the Lindeberg-Feller central limit theorem shows that a triangular array of independent random variables satisfying Lindeberg's condition is asymptotically normal after suitable scaling.

Theorem 5.1 (Lindeberg-Feller)**.** *For each $n \in \mathbb{N}$, let $Z_{n1}, Z_{n2}, \ldots, Z_{nr_n}$ be a sequence of random variables with $\mathbb{E}(Z_{ni}) = 0$ and $\text{var}(Z_{ni}) = \sigma_{ni}^2$ for $i = 1, \ldots, r_n$. Let $s_n^2 = \sum_{i=1}^{r_n} \sigma_{ni}^2$ and suppose that $r_n \to \infty$ as $n \to \infty$. Suppose the triangular array of random variables satisfies Lindeberg's condition (Definition 5.1). Then*

$$\frac{1}{s_n} \sum_{i=1}^{r_n} Z_{ni} \xrightarrow{d} N(0, 1)$$

For a proof see Loeve (1977). It can be difficult to show Lindeberg's condition directly. A stronger condition that implies the Lindeberg condition is the Lyapunov condition.

Definition 5.2 (Lyapunov's condition)**.** *For each $n \in \mathbb{N}$, let $Z_{n1}, Z_{n2}, \ldots, Z_{nr_n}$ be a sequence of random variables with $\mathbb{E}(Z_{ni}) = 0$ and $\text{var}(Z_{ni}) = \sigma_{ni}^2$ for $i = 1, \ldots, r_n$. Let $s_n^2 = \sum_{i=1}^{r_n} \sigma_{ni}^2$ and suppose that $r_n \to \infty$ as $n \to \infty$. The triangular array of random variables is said to satisfy Lyapunov's condition if there exists a $\delta > 0$ such that*

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{r_n} \mathbb{E}(|Z_{ni}|^{2+\delta}) = 0. \tag{5.15}$$

The Lyapunov condition implies the Lindeberg condition. We state this in a Lemma for later reference.

Lemma 5.1. *The Lyapunov condition implies the Lindeberg condition.*

To see this assume the Lyapunov condition is satisfied and fix $\eta > 0$. Now $|Z_{ni}| \geq \eta s_n$ implies that $1 \leq |Z_{ni}/(\eta s_n)|^\delta$. We can then form an upper bound on the sequence of partial sums that appear in Lindeberg's condition.

$$\begin{aligned}
\frac{1}{s_n^2} \sum_{i=1}^{r_n} \mathbb{E}(Z_{ni}^2 \mathbb{1}_{\{|Z_{ni}|>\eta s_n\}}) &\leq \frac{1}{s_n^2} \sum_{i=1}^{r_n} \mathbb{E}(Z_{ni}^2 |Z_{ni}/(\eta s_n)|^\delta \mathbb{1}_{\{|Z_{ni}|>\eta s_n\}}) \\
&= \frac{1}{s_n^2} \sum_{i=1}^{r_n} \mathbb{E}(|Z_{ni}|^2 |Z_{ni}/(\eta s_n)|^\delta \mathbb{1}_{\{|Z_{ni}|>\eta s_n\}}) \\
&= \frac{1}{s_n^2} \frac{1}{(\eta s_n)^\delta} \sum_{i=1}^{r_n} \mathbb{E}(|Z_{ni}|^{2+\delta} \mathbb{1}_{\{|Z_{ni}|>\eta s_n\}}) \\
&= \frac{1}{\eta^\delta} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{r_n} \mathbb{E}(|Z_{ni}|^{2+\delta}).
\end{aligned}$$

Assuming that Lyapunov's condition holds we can establish zero as an upper bound

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^{r_n} \mathbb{E}(Z_{ni}^2 \mathbb{1}_{\{|Z_{ni}|>\eta s_n\}}) &\leq \lim_{n\to\infty} \frac{1}{\eta^\delta} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{r_n} \mathbb{E}(|Z_{ni}|^{2+\delta}) \\
&= \frac{1}{\eta^\delta} \lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{r_n} \mathbb{E}(|Z_{ni}|^{2+\delta}) \\
&= 0.
\end{aligned}$$

We also have the lower bound

$$0 \leq \lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^{r_n} \mathbb{E}(Z_{ni}^2 \mathbb{1}_{\{|Z_{ni}|>\eta s_n\}}).$$

By the squeeze theorem we then have that the Lyapunov condition implies the Lindeberg condition.

We now present a useful Lemma for showing the Lyapunov condition. The result is from Billingsley (1999) and applies to triangular arrays of uniformly bounded random variables.

Lemma 5.2 (Billingsley, 1999)**.** *For each $n \in \mathbb{N}$, let $Z_{n1}, Z_{n2}, \ldots, Z_{nr_n}$ be a sequence of random variables with $\mathbb{E}(Z_{ni}) = 0$ and $\text{var}(Z_{ni}) = \sigma_{ni}^2$ for $i = 1, \ldots, r_n$. Let $s_n^2 = \sum_{i=1}^{r_n} \sigma_{ni}^2$ and suppose that $r_n \to \infty$ as $n \to \infty$. Suppose that we can form a sequence of upper bounds $(K_n)_{n\in\mathbb{N}}$ such that*

$$|Z_{ni}| \leq K_n \text{ almost surely for } i = 1, \ldots, r_n.$$

Then if $K_n/s_n \to 0$ as $n \to \infty$ the Lyapunov condition holds for the triangular array of random variables.

Lemma 5.2 is useful as it does not impose a constant uniform bound on the random variables. In the special case where $|Z_{ni}| \leq M$ almost surely for some constant M for all $n \in \mathbb{N}$ and all $i = 1, \ldots, r_n$ we have that Lyapunov's condition is satisfied providing that $s_n \to \infty$. Lemma 5.2 allows for the bound K_n to increase with n as long as the rate of growth is slower than the rate of growth of s_n. Lyapunov's condition holds providing that $K_n = o(s_n)$.

The proof of Lemma 5.2 is given below. Again fix some $\delta > 0$. If $|Z_{ni}| \leq K_n$ almost surely for $i = 1, \ldots, r_n$ it must hold that $|Z_{ni}|^\delta \leq K_n^\delta$ as $|Z_{ni}|, K_n$ and δ are all positive. As such $|Z_{ni}|^{2+\delta} = |Z_{ni}|^2|Z_{ni}|^\delta \leq |Z_{ni}|^2 K_n^\delta$. We can then form an upper bound on the sequence of partial sums that appear in Lyapunov's condition.

$$\begin{aligned}\frac{1}{s_n^{2+\delta}} \sum_{i=1}^{r_n} \mathbb{E}(|Z_{ni}|^{2+\delta}) &\leq \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{r_n} \mathbb{E}(|Z_{ni}|^2) K_n^\delta \\ &= \frac{K_n^\delta}{s_n^{2+\delta}} \sum_{i=1}^{r_n} \mathbb{E}|Z_{ni}|^2. \\ &= \frac{K_n^\delta}{s_n^{2+\delta}} s_n^2 \\ &= \left(\frac{K_n}{s_n}\right)^\delta. \end{aligned} \tag{5.16}$$

Now assuming that $K_n = o(s_n)$ we have that $K_n/s_n \to 0$ as $n \to \infty$. We then also have that

$$\begin{aligned}\lim_{n\to\infty} \left(\frac{K_n}{s_n}\right)^\delta &= \left(\lim_{n\to\infty} \frac{K_n}{s_n}\right)^\delta \\ &= 0,\end{aligned}$$

as the exponentiation by $\delta > 0$ is a continuous function. Now taking limits on both sides of the inequality (5.16):

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{r_n} \mathbb{E}(|Z_{ni}|^{2+\delta}) \leq \lim_{n\to\infty} \left(\frac{K_n}{s_n}\right)^\delta \tag{5.17}$$

$$= 0. \tag{5.18}$$

We also have the lower bound

$$0 \leq \lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{r_n} \mathbb{E}(|Z_{ni}|^{2+\delta}).$$

By the squeeze theorem we then have that $K_n = o(s_n)$ is sufficient for Lyapunov's condition to hold.

The triangular array of independent random variables in Theorem 4.4 satisfies Lyapunov's condition by Lemma 5.2. As the Lyapunov condition implies the Lindeberg condition (Lemma 5.1) the general Lindeberg-Feller central limit theorem (Theorem 5.1) gives asymptotic normality of the scaled row sums, thus proving Theorem 4.4.

5.6 Proof of Theorem 4.3 (Sketching central limit theorem)

Assumption 1 Let the singular value decomposition of the $n \times d$ source dataset be given by $\boldsymbol{A}_{(n)} = \boldsymbol{U}_{(n)}\boldsymbol{D}_{(n)}\boldsymbol{V}_{(n)}^\mathsf{T}$. Let $\boldsymbol{u}_{(n)i}^\mathsf{T}$ give the ith row in $\boldsymbol{U}_{(n)}$. Assume that the maximum leverage score tends to zero, that is

$$\lim_{n\to\infty} \max_{i=1,\ldots,n} \|\boldsymbol{u}_{(n)i}\|_2^2 = 0.$$

Theorem 4.3 gives the sketching central limit theorem.

Theorem. *Consider a fixed sequence of arbitrary $n \times d$ data matrices $\boldsymbol{A}_{(n)}$, where d is fixed. Let $\boldsymbol{A}_{(n)} = \boldsymbol{U}_{(n)}\boldsymbol{D}_{(n)}\boldsymbol{V}_{(n)}^\mathsf{T}$ represent the singular value decomposition of $\boldsymbol{A}_{(n)}$. Let $\boldsymbol{S}$ be a $k \times n$ Hadamard or Clarkson-Woodruff sketching matrix where k is also fixed. Suppose that Assumption 1 on the maximum leverage score is satisfied. Then as n tends to infinity with k and d fixed,*

$$[\widetilde{\boldsymbol{A}}\boldsymbol{V}_{(n)}\boldsymbol{D}_{(n)}^{-1} \mid \boldsymbol{A}_{(n)}] \xrightarrow{d} \mathrm{MN}(\boldsymbol{0}, \boldsymbol{I}_k, \boldsymbol{I}_d/k).$$

To prove the sketching central limit theorem it helps to restate Lemma 5.2. This helps to show the importance of the leverage scores in establishing asymptotic normality. Lemma 5.2 provided a sufficient condition for showing that Lindeberg's condition holds. We can restate Lemma 5.2 in terms of a normalised triangular array.

Theorem 5.2 (Billingsley, 1999)**.** *For each $n \in \mathbb{N}$ let $Z_{n1}, Z_{n2}, \ldots, Z_{nr_n}$ be a sequence of random variables with $\mathbb{E}[Z_{ni}] = 0$ and $var(Z_{ni}^2) = \sigma_{ni}^2$ for $i = 1, \ldots, r_n$. Define $s_n^2 = \sum_{i=1}^{r_n} \sigma_{ni}^2$ each n. Suppose that the rows of the triangular array are standardised such that $s_n^2 = 1$ for all n. Suppose that $r_n \to \infty$ as $n \to \infty$. Suppose we have a sequence of upper bounds (Z_n) such that $|Z_{ni}| \leq K_n$ almost surely for all $i = 1, \ldots, r_n$. Then a sufficient condition for Lyapnuov's condition to hold is $K_n \to 0$ as $n \to \infty$.*

The standardisation of the triangular array gives an intuitive condition for Lyapunov's and hence Lindeberg's condition to hold. We require that $K_n \to 0$ as $n \to \infty$. We require that the upper bound tends to zero. All the random variables in the row must converge almost surely to zero. Almost sure convergence is stronger than convergence in probability and rules out pathological cases where a single random variable in a row can take a large value with small probability. Assumption 1 on the leverage scores in the sketching central limit theorem enforces a bounded growth condition that relates to Theorem 5.2.

Let $n \in \mathbb{N}$ index the sequence of source datasets of increasing size. We assume that the source dataset consists of r_n observations where $r_n \to \infty$ as $n \to \infty$. For now we can take take $r_n = n$ to ease interpretation. We take the singular value decomposition of each dataset $\boldsymbol{A}_{(n)} = \boldsymbol{U}_{(n)}\boldsymbol{D}_{(n)}\boldsymbol{V}_{(n)}^\mathsf{T}$. All results in this section treat the source dataset $\boldsymbol{A}_{(n)}$ as fixed, only the sketching matrix is random. We consider the sequence of whitened sketched datasets

$$\begin{aligned}\widetilde{\boldsymbol{A}}\boldsymbol{V}_{(n)}\boldsymbol{D}_{(n)}^{-1} &= (\boldsymbol{S}\boldsymbol{A})\boldsymbol{V}_{(n)}\boldsymbol{D}_{(n)}^{-1} \\ &= \boldsymbol{S}\boldsymbol{U}_{(n)}\boldsymbol{D}_{(n)}\boldsymbol{V}_{(n)}^\mathsf{T}\boldsymbol{V}_{(n)}\boldsymbol{D}_{(n)}^{-1} \\ &= \boldsymbol{S}\boldsymbol{U}_{(n)}.\end{aligned}$$

The whitened sketched dataset $\widetilde{\boldsymbol{A}}\boldsymbol{V}_{(n)}\boldsymbol{D}_{(n)}^{-1}$ has a $MN(\boldsymbol{0}, \boldsymbol{I}_k, \boldsymbol{I}_d/k)$ distribution when $\boldsymbol{S}$ is a Gaussian sketch. We need to show that as n tends to infinity, $\boldsymbol{S}\boldsymbol{U}_{(n)}$ convergences in distribution to a $MN(\boldsymbol{0}, \boldsymbol{I}_k, \boldsymbol{I}_d/k)$ random matrix for both the Clarkson-Woodruff and Hadamard sketches.

Let $\boldsymbol{u}_{(n)i}^\mathsf{T}$ denote row i of the matrix of left singular vectors $\boldsymbol{U}_{(n)}$. We write $\boldsymbol{u}_{(n)i}^\mathsf{T}$ so that that we can form a triangular array of left singular vectors. Taking $r_n = n$, the first three rows of the triangular array can be written as

$$\begin{array}{lll}\boldsymbol{u}_{(1)1} & & \\ \boldsymbol{u}_{(2)1} & \boldsymbol{u}_{(2)2} & \\ \boldsymbol{u}_{(3)1} & \boldsymbol{u}_{(3)2} & \boldsymbol{u}_{(3)3}\end{array}$$

An important property is that for all n, the sum of the norms of the leverage scores always equals the number of variables in the source dataset d.

$$\sum_{i=1}^{r_n} \|\boldsymbol{u}_{(n)i}\|_2^2 = d. \tag{5.19}$$

As n increases, the typical norm of each vector $\boldsymbol{u}_{(n)i}$, $i \in \{1, \ldots, r_n\}$ is expected to decrease. For completeness we restate Assumption 1 in terms of the triangular array formulation.

Assumption 1 Let the singular value decomposition of the $r_n \times d$ source dataset be given by $\boldsymbol{A}_{(n)} = \boldsymbol{U}_{(n)}\boldsymbol{D}_{(n)}\boldsymbol{V}_{(n)}^\mathsf{T}$. Let $\boldsymbol{u}_{(n)i}^\mathsf{T}$ give the ith row in $\boldsymbol{U}_{(n)}$ for $i = 1, \ldots, r_n$. Assume that the maximum leverage score tends to zero, that is

$$\lim_{n\to\infty} \max_{i=1,\ldots,r_n} \|\boldsymbol{u}_{(n)i}\|_2^2 = 0.$$

This increasing collection of smaller quantities is similar to the behaviour of the triangular array of random variables in Theorem 5.2. The standardisation property in equation (5.19), namely that $\sum_{i=1}^{r_n} \|\boldsymbol{u}_{(n)i}\|_2^2 = d$ for all n is similar to the assumption that $s_n = 1$ in each row of the triangular array of random variables in Theorem 5.2. Assumption 1 on the leverage scores, where the maximum individual norm tends to zero is similar to the assumption that $K_n \to 0$ in Theorem 5.2. This will be made more explicit in the proofs. Before moving on we make a note that assumption 1 also implies that the maximum square root of the leverage scores also tends to zero. As

$$\max_{i=1,\ldots,r_n} \|\boldsymbol{u}_{(n)i}\|_2 = \left(\max_{i=1,\ldots,r_n} \|\boldsymbol{u}_{(n)i}\|_2^2 \right)^{1/2} \tag{5.20}$$

We have that

$$\begin{aligned} \lim_{n\to\infty} \max_{i=1,\ldots,r_n} \|\boldsymbol{u}_{(n)i}\|_2 &= \lim_{n\to\infty} \left(\max_{i=1,\ldots,r_n} \|\boldsymbol{u}_{(n)i}\|_2^2 \right)^{1/2} \\ &= \left(\lim_{n\to\infty} \max_{i=1,\ldots,r_n} \|\boldsymbol{u}_{(n)i}\|_2^2 \right)^{1/2} \\ &= 0. \end{aligned} \tag{5.21}$$

To establish joint asymptotic normality of the sketched data matrix we use the Cramér-Wold device.

Lemma 5.3 (Cramér-Wold device)**.** *Let $\boldsymbol{z}$ be a real valued random vector of dimension v and let $(\boldsymbol{z}_n)$ be a sequence of real valued random vectors of the same fixed dimension v. The sequence of random vectors $(\boldsymbol{z}_n)$ converges in distribution to $\boldsymbol{z}$ as n tends to infinity if and only if $(\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{z}_n)$ converges in distribution to $\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{z}$ for all unit vectors $\boldsymbol{\lambda} \in \mathbb{R}^v$.*

Let $\boldsymbol{z}_n$ represent the kd length vector formed by stacking transposed rows of the whitened sketched dataset $\widetilde{\boldsymbol{U}} = \boldsymbol{S}\boldsymbol{U}_{(n)}$. Let $\widetilde{\boldsymbol{u}}_j^\mathsf{T}$ give row j in $\widetilde{\boldsymbol{U}}$ for $j = 1, \ldots, k$. Formally,

$$\boldsymbol{z}_n = \begin{bmatrix} \widetilde{\boldsymbol{u}}_1 \\ \widetilde{\boldsymbol{u}}_2 \\ \vdots \\ \widetilde{\boldsymbol{u}}_k \end{bmatrix}. \tag{5.22}$$

Let us define the random matrix $k \times d$ random matrix $\boldsymbol{W}$ as having the matrix normal distribution

$$\boldsymbol{W} \sim MN(\boldsymbol{0}, \boldsymbol{I}_k, \boldsymbol{I}_d/k)$$

Let $\boldsymbol{w}_i^\mathsf{T}$ refer to row i in $\boldsymbol{W}$ for $i = 1, \ldots, k$. Let $\boldsymbol{z}_L$ refer to the stacked transposed rows of $\boldsymbol{W}$, so

$$\boldsymbol{z}_L = \begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \\ \vdots \\ \boldsymbol{w}_k \end{bmatrix}. \tag{5.23}$$

Let $\boldsymbol{\lambda}$ be an arbitrary unit vector in $\mathbb{R}^{k\times d}$. It will be useful to also partition the vector $\boldsymbol{\lambda}$ into k sub-vectors,

$$\boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_2 \\ \vdots \\ \boldsymbol{\lambda}_k \end{bmatrix}, \tag{5.24}$$

where $\boldsymbol{\lambda}_j$ is a d-dimensional vector for $j = 1, \ldots, k$. For any unit vector $\boldsymbol{\lambda} \in \mathbb{R}^{k \times d}$, $\boldsymbol{\lambda}^\mathsf{T} \boldsymbol{z}_L$ is distributed as $N(0, 1/k)$. We will aim to show that the distribution of the whitened sketched data $\boldsymbol{S}\boldsymbol{A}_{(n)}\boldsymbol{V}_{(n)}\boldsymbol{D}_{(n)}^{-1}$ converges to that of $\boldsymbol{W}$ through the Cramér-Wold device. We must show that for any fixed $k \times d$ length unit vector $\boldsymbol{\lambda}$, $\boldsymbol{\lambda}^\mathsf{T} \boldsymbol{z}_n$ converges in distribution to $N(0, 1/k)$ as $n \to \infty$.

We will rely on a central limit theorem for jointly symmetric, pairwise independent random variables (Pruss and Szynal, 2000). A collection of random variables $(Z_1, \ldots Z_n)$ is said to be jointly symmetric if $(Z_1, \ldots Z_n)$ has the same distribution as $(q_1 Z_1, \ldots, q_n Z_n)$, where $q_i \in \{+1, -1\}$ for $i = 1, \ldots, n$. Given a set of random variables $Y_1, \ldots, Y_n$, a jointly symmetric collection $Z_1, \ldots, Z_n$ can be formed by sampling n independent Rademacher random variables $h_1, \ldots, h_n$, and setting $Z_i = h_i Y_i$ (Pruss and Szynal, 2000). It is possible to establish a central limit theorem for jointly symmetric, pairwise independent random variables.

Theorem 5.3 (Pruss and Szynal (2000), Theorem 1, Corollary 2)**.** *For each $n \in \mathbb{N}$, let $Z_{n1}, Z_{n2}, \ldots, Z_{nr_n}$ be a sequence of jointly symmetric pairwise independent random variables with $\mathbb{E}(Z_{ni}) = 0$ and $\operatorname{var}(Z_{ni}) = \sigma_{ni}^2$ for $i = 1, \ldots, r_n$. Let $s_n^2 = \sum_{i=1}^{r_n} \sigma_{ni}^2$ and assume that $r_n \to \infty$ as $n \to \infty$. Suppose the triangular array of random variables satisfies Lindeberg's condition. Then as $n \to \infty$, $s_n^{-1} \sum_{i=1}^{r_n} Z_{ni}$ converges in distribution to $N(0, 1)$.*

Not all triangular arrays with pairwise independent random variables in each row satisfy a central limit theorem. The joint symmetry property is very important (Pruss and Szynal, 2000; Janson, 1988).

To use Theorem 5.3 we need to show that the triangular array of random variables satisfies Lindeberg's condition. As discussed this can be very difficult to establish directly. If the triagular array of random variables can be appropriately bounded, we can use Theorem 5.2 to show that Lyapunov's condition holds, and subsequently that Lindeberg's condition holds.

This is the approach we take in proving the sketching central limit theorem. The Cramér-Wold device is used to reduce the study of multivariate convergence to univariate convergence. We can then form a triangular array of random variables such that elements in each row are jointly symmetric and pairwise independent. We then show that triangular array satisfies Lindeberg's condition using Theorem 5.2. Assumption 1 on the maximum leverage score enforces the necessary cap on the rate of growth. Theorem 5.3 is then used to establish asymptotic normality.

5.6.1 Clarkson-Woodruff sketch

The Clarkson-Woodruff sketch can be represented as the product of two independent random matrices, $\boldsymbol{S} = \boldsymbol{\Gamma}\boldsymbol{D}$, where $\boldsymbol{\Gamma}$ is a random $k \times n$ matrix and $\boldsymbol{D}$ is a random $n \times n$ matrix. The diagonal matrix $\boldsymbol{D}$ contains n independent Rademacher random variables on the diagonal. Let $h_i \in \{+1, -1\}$ be the random sign in element D_{ii}. The matrix $\boldsymbol{\Gamma}$ is formed by choosing one element in each column independently and setting the entry to $+1$. Element Γ_{ij} is equal to $+1$ if we add observation i in the original dataset to sketched observation j. The signs in row i are flipped if h_i is equal to negative one. Each observation in the original dataset is assigned to one sketched observation as each column of $\boldsymbol{\Gamma}$ contains a single $+1$ entry. Using a Clarkson-Woodruff sketch row j in the sketched data matrix can be represented as

$$\widetilde{\boldsymbol{u}}_j^\mathsf{T} = \sum_{i=1}^{n} h_i \Gamma_{ij} \boldsymbol{u}_{(n)i}^\mathsf{T},$$

where h_i represents the random sign flip applied to row i of the original data matrix, and Γ_{ij} is the indicator variable which is equal to one if row i of the original data is added to row j of the sketched dataset.

Let us consider the linear combination $\boldsymbol{\lambda}^\mathsf{T} \boldsymbol{z}$, where $\boldsymbol{\lambda}$ and $\boldsymbol{z}$ are defined as in (5.22) and (5.24) respectively. The sum over the k rows in the sketched dataset can be rearranged into a sum over the n

rows in the source dataset,

$$\begin{aligned}\boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{z}_n &= \sum_{j=1}^{k} \boldsymbol{\lambda}_j^{\mathsf{T}} \widetilde{\boldsymbol{u}}_j \\ &= \sum_{j=1}^{k} \boldsymbol{\lambda}_j^{\mathsf{T}} \sum_{i=1}^{n} h_i \Gamma_{ij} \boldsymbol{u}_{(n)i} \\ &= \sum_{i=1}^{n} h_i \sum_{j=1}^{k} \Gamma_{ij} \boldsymbol{\lambda}_j^{\mathsf{T}} \boldsymbol{u}_{(n)i}. \end{aligned} \tag{5.25}$$

The scalar $\boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{z}_n$ is equal to the sum of n independent random variables. Independence holds as the signs flips h_i on each observation are independent, and each column of $\boldsymbol{\Gamma}$ is independent.

In the language of Theorem 5.2 we can form a triangular array of random variables setting

$$Z_{ni} = h_i \sum_{j=1}^{k} \Gamma_{ij} \boldsymbol{\lambda}_j^{\mathsf{T}} \boldsymbol{u}_{(n)i}. \tag{5.26}$$

for $i = 1, \ldots, n$ and $n \in \mathbb{N}$. The linear combination in (5.25) then be expressed as a row sum over the triangular array defined in (5.26):

$$\boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{z}_n = \sum_{i=1}^{n} Z_{ni}. \tag{5.27}$$

Our goal of showing that $\boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{z}_n$ converges in distribution to a $N(0, 1/k)$ random variable is achieved if we can show that $\sum_{i=1}^{n} Z_{ni}$ converges in distribution to a $N(0, 1/k)$ random variable.

It is worth making a connection to Theorem 5.3, because of the random sign flips h_i appearing in (5.26), we have a sequence of mutually independent jointly symmetric random variables. Mutually independent random variables are also necessarily pairwise independent. Theorem 5.3 can be used to establish asymptotic normality of the sum in (5.27) and hence the linear combination $\boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{z}_n$. To show that the triangular array of random variables defined in (5.26) satisfies Lindeberg's condition we use Theorem 5.2. Set $s_n^2 = \sum_{i=1}^{n} \operatorname{var}(Z_{ni})$. We first determine s_n^2. We then form the necessary sequence of upper bounds K_n such that $|Z_{ni}| \leq K_n$ almost surely for $i = 1, \ldots, n$. The variance of a single term in the sum (5.25) is

$$\operatorname{var}(Z_{ni}) = \operatorname{var}\left(h_i \sum_{j=1}^{k} \Gamma_{ij} \boldsymbol{\lambda}_j^{\mathsf{T}} \boldsymbol{u}_{(n)i} \right) \tag{5.28}$$

$$= \sum_{j=1}^{k} \frac{1}{k} \boldsymbol{\lambda}_j^{\mathsf{T}} \boldsymbol{u}_{(n)i} \boldsymbol{u}_{(n)i}^{\mathsf{T}} \boldsymbol{\lambda}_j. \tag{5.29}$$

The row-wise variance totals s_n^2 are then

$$\begin{aligned}
s_n^2 &= \sum_{i=1}^{n} \operatorname{var}(Z_{ni}) \\
&= \sum_{i=1}^{n} \operatorname{var}\left(h_i \sum_{j=1}^{k} \Gamma_{ij} \boldsymbol{\lambda}_j^{\mathsf{T}} \boldsymbol{u}_{(n)i} \right) \\
&= \frac{1}{k} \sum_{i=1}^{n} \sum_{j=1}^{k} \boldsymbol{\lambda}_j^{\mathsf{T}} \boldsymbol{u}_{(n)i} \boldsymbol{u}_{(n)i}^{\mathsf{T}} \boldsymbol{\lambda}_j \\
&= \frac{1}{k} \sum_{j=1}^{k} \boldsymbol{\lambda}_j^{\mathsf{T}} \left(\sum_{i=1}^{n} \boldsymbol{u}_{(n)i} \boldsymbol{u}_{(n)i}^{\mathsf{T}} \right) \boldsymbol{\lambda}_j \\
&= \frac{1}{k} \sum_{j=1}^{k} \boldsymbol{\lambda}_j^{\mathsf{T}} \boldsymbol{U}_{(n)}^{\mathsf{T}} \boldsymbol{U}_{(n)} \boldsymbol{\lambda}_j \\
&= \frac{1}{k} \sum_{j=1}^{k} \boldsymbol{\lambda}_j^{\mathsf{T}} \boldsymbol{I}_d \boldsymbol{\lambda}_j. \\
&= \frac{1}{k} \sum_{j=1}^{k} \boldsymbol{\lambda}_j^{\mathsf{T}} \boldsymbol{\lambda}_j \\
&= \frac{1}{k}.
\end{aligned}$$

The fact that $\boldsymbol{U}_{(n)}^{\mathsf{T}} \boldsymbol{U}_{(n)} = \boldsymbol{I}_d$ for all n serves as a useful normalisation to give stable limiting behaviour. The step in the last line follows as we have taken $\boldsymbol{\lambda}$ to be a unit vector. We have $s_n = 1/k$ for all n in the triangular array. We now establish a sequence of upper bounds (K_n). As the random variables in the construction of construction of the sketch are bounded, we can bound the random variables in the triangular array using the leverage scores of the sequence of source dataset. Now as the random sign $h_i \in \{+1, -1\}$

$$\begin{aligned}
|Z_{ni}| &= |h_i \sum_{j=1}^{k} \Gamma_{ij} \boldsymbol{\lambda}_j^{\mathsf{T}} \boldsymbol{u}_{(n)i}| \\
&= |\left(\sum_{j=1}^{k} \Gamma_{ij} \boldsymbol{\lambda}_j^{\mathsf{T}} \right) \boldsymbol{u}_{(n)i}|. \qquad (5.30)
\end{aligned}$$

Now by the Cauchy-Schwarz inequality

$$|\left(\sum_{j=1}^{k} \Gamma_{ij} \boldsymbol{\lambda}_j^{\mathsf{T}} \right) \boldsymbol{u}_{(n)i}| \leq \|\sum_{j=1}^{k} \Gamma_{ij} \boldsymbol{\lambda}_j\|_2 \|\boldsymbol{u}_{(n)i}\|_2 \qquad (5.31)$$

Now as $\Gamma_{ij} = 1$ for a single $j \in \{1, \ldots, k\}$ and is zero otherwise we have that

$$\begin{aligned}
\|\sum_{j=1}^{k} \Gamma_{ij} \boldsymbol{\lambda}_j\|_2 &\leq \max_{\mathrm{j}=1,\,\ldots,\,\mathrm{k}} \|\lambda_j\|_2 \\
&\leq 1. \qquad (5.32)
\end{aligned}$$

The last line follows as we have taken $\boldsymbol{\lambda}$ to be a unit vector. Substituting (5.32) and (5.31) into (5.30) we arrive at

$$|Z_{ni}| \leq \|\boldsymbol{u}_{(n)i}\|_2.$$

We can then form the sequence of upper bounds K_n,

$$K_n = \max_{i=1,\ldots,n} \|\boldsymbol{u}_{(n)i}\|_2.$$

We have that $|Z_{ni}| \leq K_n$ almost surely for $i = 1, \ldots, n$ and $n \in \mathbb{N}$. Assumption 1 controls the limiting behaviour of $K_n = \max_{i=1,\ldots,n} \|\boldsymbol{u}_{(n)i}\|_2$ (recall equation (5.21)). Taking limits and using Assumption 1 shows that $K_n \to 0$,

$$\begin{aligned}\lim_{n\to\infty} K_n &= \lim_{n\to\infty} \max_{i=1,\ldots,n} \|\boldsymbol{u}_{(n)i}\|_2 \\ &= 0.\end{aligned}$$

By theorem 5.2 we have that the triangular array of random variables in (5.26) satisfies Lindeberg's condition. As such the conditions of Theorem 5.3 are satisfied, giving that $\boldsymbol{\lambda}^\intercal \boldsymbol{z}_n$ converges in distribution to $N(0, 1/k)$. Finally, the Cramér-Wold device gives that the whitened sketched dataset has a limiting matrix normal distribution, that is $\widetilde{\boldsymbol{A}}\boldsymbol{V}_{(n)}\boldsymbol{D}_{(n)}^{-1}$ converges in distribution to a $MN\left(\boldsymbol{0}_{k\times d}, \boldsymbol{I}_k, \boldsymbol{I}_d/k\right)$ random matrix.

5.6.2 Hadamard sketch

Recall that the Hadamard sketch is defined through $\boldsymbol{S} = \boldsymbol{\Phi}\boldsymbol{H}\boldsymbol{D}/\sqrt{k}$. Here $\boldsymbol{H}$ is a Hadamard matrix. Hadamard matrices are square matrices with 2^n rows for some integer n. To take limits we have to define our sequence of source datasets ($\boldsymbol{A}_{(n)} = \boldsymbol{U}_{(n)}\boldsymbol{D}_{(n)}\boldsymbol{V}_{(n)}^\intercal$) as having $r_n = 2^n$ rows for $n \in \mathbb{N}^+$. In practice when taking a Hadamard sketch we pad the original dataset with zeros if the original number of observations is not a power of two. To rigourously establish asymptotic normality for the Hadamard sketch we have to take $r_n = 2^n$. The first three rows of the triangular array of left singular vectors now looks like

$$\begin{array}{cccc}\boldsymbol{u}_{(1)1} & & & \\ \boldsymbol{u}_{(2)1} & \boldsymbol{u}_{(2)2} & & \\ \boldsymbol{u}_{(3)1} & \boldsymbol{u}_{(3)2} & \boldsymbol{u}_{(3)3} & \boldsymbol{u}_{(n)4}.\end{array}$$

The intuition is the same as with the Clarkson-Woodruff sketch, as we move down the rows n we expect the norms $\boldsymbol{u}_{(n)i}$, $i \in \{1, \ldots, 2^n\}$ to decrease. This follows from the implicit row-wise normalisation property

$$\sum_{i=1}^{r_n} \|\boldsymbol{u}_{(n)i}\|_2^2 = d.$$

The indexing change to $r_n = 2^n$ instead of $r_n = n$ has very little impact on the underlying arguments.

There are two independent sources of randomness in a Hadamard sketch, the $r_n = 2^n$ independent random Rademacher variables in the diagonal matrix $\boldsymbol{D}$, and the random matrix $\boldsymbol{\Phi}$ which subsamples k rows with replacement from the Hadamard matrix $\boldsymbol{H}$. Hadamard matrices have a number of properties that we will use (Anderson, 1997, section 3.2).

- (P1) The first column contains all ones.
- (P2) Every column other than the first contains an equal number of $+1$ and -1 entries.
- (P3) Consider any two different columns i and s, where $i, s \in \{2, \ldots, r_n\}$, $i \neq s$. Columns i and s will have $+1$ together in a quarter of the rows, and -1 together in a quarter of the rows. Furthermore, a quarter of the rows will have $+1$ in column i and -1 in column s. Similarly, a quarter of the rows will have -1 in column i and $+1$ in column s.

Let $\boldsymbol{M}$ represent the random $k \times n$ matrix from the subsampling operation $\boldsymbol{M} = \Phi\boldsymbol{H}$. Let m_{ji} refer to the element in row j and column i of $\boldsymbol{M}$. Each element in $\boldsymbol{M}$ is equal to $+1$ or -1. Let $h_i \in \{+1, -1\}$ be the random sign in element D_{ii}. We now represent the Hadamard sketch as $\boldsymbol{S} = \boldsymbol{M}\boldsymbol{D}/\sqrt{k}$.

The structure of the Hadamard matrix gives the random matrix $\boldsymbol{M}$ some useful properties. Consider an arbitrary row j in $\boldsymbol{M}$. By (P1) listed above regarding the first column of $\boldsymbol{M}$, $m_{j1} = 1$ with probability one. For the other columns, $m_{ji} = 1$ with probability half, and $m_{ji} = -1$ with probability half for $i = 2, \ldots, r_n$ by (P2). By (P3) listed above, we have pairwise independence between elements in row j of $\boldsymbol{M}$, that is $p(m_{ji}|m_{js}) = p(m_{ji})$ for $i, s \in \{1, \ldots, r_n\}$, $i \neq s$. As rows of $\boldsymbol{M}$ are sampled independently, each column of $\boldsymbol{M}$ is pairwise independent.

Row j in the sketched dataset is given by

$$\widetilde{\boldsymbol{u}}_j^{\mathsf{T}} = \frac{1}{\sqrt{k}} \sum_{i=1}^{r_n} m_{ji} h_i \boldsymbol{u}_{(n)i}^{\mathsf{T}}.$$

Let us again consider the linear combination $\boldsymbol{\lambda}^{\mathsf{T}} \boldsymbol{z}_n$, where $\boldsymbol{\lambda}$ and $\boldsymbol{z}_n$ are defined as in (5.22) and (5.24) respectively. The sum over the k rows in the sketched dataset can be rearranged into a sum over the $r_n = 2^n$ rows in the source dataset,

$$\begin{aligned} \boldsymbol{\lambda}^{\mathsf{T}} \boldsymbol{z}_n &= \sum_{j=1}^{k} \boldsymbol{\lambda}_j^{\mathsf{T}} \widetilde{\boldsymbol{u}}_j \\ &= \frac{1}{\sqrt{k}} \sum_{j=1}^{k} \boldsymbol{\lambda}_j^{\mathsf{T}} \sum_{i=1}^{r_n} m_{ji} h_i \boldsymbol{u}_{(n)i} \\ &= \frac{1}{\sqrt{k}} \sum_{i=1}^{r_n} h_i \left(\sum_{j=1}^{k} m_{ji} \boldsymbol{\lambda}_j^{\mathsf{T}} \right) \boldsymbol{u}_{(n)i}. \end{aligned} \tag{5.33}$$

In the language of Theorem 5.2 we can form a triangular array of random variables setting

$$Z_{ni} = \frac{1}{\sqrt{k}} h_i \left(\sum_{j=1}^{k} m_{ji} \boldsymbol{\lambda}_j^{\mathsf{T}} \right) \boldsymbol{u}_{(n)i}. \tag{5.34}$$

for $i = 1, \ldots, r_n$ and $n \in \mathbb{N}$. The linear combination in (5.33) can then be expressed as a row sum of the triangular array defined by (5.34)

$$\boldsymbol{\lambda}^{\mathsf{T}} \boldsymbol{z}_n = \sum_{i=1}^{r_n} Z_{ni}. \tag{5.35}$$

Our goal of showing that $\boldsymbol{\lambda}^{\mathsf{T}} \boldsymbol{z}_n$ converges in distribution to a $N(0, 1/k)$ random variable is achieved if we can show that $\sum_{i=1}^{n} Z_{ni}$ converges in distribution to a $N(0, 1/k)$ random variable.

The sequence of random variables in each row of the triangular array $Z_{n1}, \ldots, Z_{nr_n}$ are not mutually independent over $i = 1 \ldots, r_n$. This is because the columns of $\boldsymbol{M}$ are not mutually independent. However, as the columns of $\boldsymbol{M}$ are pairwise independent, the random sums $\sum_{j=1}^{k} m_{ji} \boldsymbol{\lambda}_j^{\mathsf{T}}$ appearing in (5.34) are also pairwise independent. Again making a connection to Theorem 5.3, the independent sign flips h_i appearing in (5.34) ensure that the random variables in each row of the triangular array are jointly symmetric and pairwise independent.

Theorem 5.3 can be used to establish asymptotic normality of the sum in (5.35) and hence the linear combination $\boldsymbol{\lambda}^{\mathsf{T}} \boldsymbol{z}_n$. To show that the triangular array of random variables defined in (5.34) satisfies Lindeberg's condition we use Theorem 5.2. Set $s_n^2 = \sum_{i=1}^{n} \operatorname{var}(Z_{ni})$. We first determine s_n^2. We then form the necessary sequence of upper bounds K_n such that $|Z_{ni}| \leq K_n$ almost surely for $i = 1, \ldots, n$.

We start by considering the variance of a single term in the triangular array $\operatorname{var}(Z_{ni})$. We have that

$$\operatorname{var}(Z_{ni}) = \frac{1}{k} \operatorname{var} \left(h_i \left(\sum_{j=1}^{k} m_{ji} \boldsymbol{\lambda}_j^{\mathsf{T}} \right) \boldsymbol{u}_{(n)i} \right) \tag{5.36}$$

It is important to consider the covariance between the elements of the sum over $j = 1, \ldots, k$. For $i \neq 1$ and $j, v \in \{1, \ldots, k\}$, $j \neq v$ the covariance is zero

$$\begin{aligned}\operatorname{cov}\left(h_i m_{ji} \boldsymbol{\lambda}_j^\mathsf{T} \boldsymbol{u}_{(n)i}, h_i m_{vi} \boldsymbol{\lambda}_v^\mathsf{T} \boldsymbol{u}_{(n)i}\right) &= \mathbb{E}\left[h_i^2 m_{ji} m_{vi} \boldsymbol{\lambda}_j^\mathsf{T} \boldsymbol{u}_{(n)i} \boldsymbol{\lambda}_v^\mathsf{T} \boldsymbol{u}_{(n)i}\right] \\ &= \mathbb{E}\left[m_{ji} m_{vi}\right] \boldsymbol{\lambda}_j^\mathsf{T} \boldsymbol{u}_{(n)i} \boldsymbol{\lambda}_v^\mathsf{T} \boldsymbol{u}_{(n)i} \\ &= 0.\end{aligned}$$

We use (P2) to conclude that $\mathbb{E}\left[m_{ji} m_{vi}\right] = 0$. Therefore for $i = 2, \ldots, r_n$

$$\begin{aligned}\operatorname{var}(Z_{ni}) &= \frac{1}{k} \operatorname{var}\left(\sum_{j=1}^{k} h_i m_{ji} \boldsymbol{\lambda}_j^\mathsf{T} \boldsymbol{u}_{(n)i}\right) \\ &= \frac{1}{k} \sum_{j=1}^{k} \operatorname{var}\left(h_i m_{ji} \boldsymbol{\lambda}_j^\mathsf{T} \boldsymbol{u}_{(n)i}\right) \\ &= \frac{1}{k} \sum_{j=1}^{k} \boldsymbol{\lambda}_j^\mathsf{T} \boldsymbol{u}_{(n)i} \boldsymbol{u}_{(n)i}^\mathsf{T} \boldsymbol{\lambda}_j. \end{aligned} \tag{5.37}$$

Results are different for $i = 1$ as the first column of the Hadamard matrix is all ones (P1). For $j, v \in \{1, \ldots, k\}$, $j \neq v$ the covariance is

$$\begin{aligned}\operatorname{cov}\left(h_1 m_{j1} \boldsymbol{\lambda}_j^\mathsf{T} \boldsymbol{u}_{(n)1}, h_1 m_{v1} \boldsymbol{\lambda}_v^\mathsf{T} \boldsymbol{u}_{(n)1}\right) &= \mathbb{E}\left[h_1^2 m_{j1} m_{v1} \boldsymbol{\lambda}_j^\mathsf{T} \boldsymbol{u}_{(n)1} \boldsymbol{\lambda}_v^\mathsf{T} \boldsymbol{u}_{(n)1}\right] \\ &= \mathbb{E}\left[m_{j1} m_{v1}\right] \boldsymbol{\lambda}_j^\mathsf{T} \boldsymbol{u}_{(n)1} \boldsymbol{\lambda}_v^\mathsf{T} \boldsymbol{u}_{(n)1} \\ &= \boldsymbol{\lambda}_j^\mathsf{T} \boldsymbol{u}_{(n)1} \boldsymbol{\lambda}_v^\mathsf{T} \boldsymbol{u}_{(n)1}.\end{aligned}$$

From (P1) $m_{j1} = m_{v1} = 1$. Now using the Cauchy-Schwarz inequality,

$$\begin{aligned}\left|\operatorname{cov}\left(h_1 m_{j1} \boldsymbol{\lambda}_j^\mathsf{T} \boldsymbol{u}_{(n)1}, h_1 m_{v1} \boldsymbol{\lambda}_v^\mathsf{T} \boldsymbol{u}_{(n)1}\right)\right| &= \left|\boldsymbol{\lambda}_j^\mathsf{T} \boldsymbol{u}_{(n)1}\right| \left|\boldsymbol{\lambda}_v^\mathsf{T} \boldsymbol{u}_{(n)1}\right| \\ &\leq \|\boldsymbol{\lambda}_j\|_2 \|\boldsymbol{u}_{(n)1}\|_2 \|\boldsymbol{\lambda}_v\|_2 \|\boldsymbol{u}_{(n)1}\|_2 \\ &\leq \|\boldsymbol{u}_{(n)1}\|_2 \|\boldsymbol{u}_{(n)1}\|_2 \\ &= \|\boldsymbol{u}_{(n)1}\|_2^2\end{aligned}$$

The second last last uses the fact that $\boldsymbol{\lambda}$ is a unit vector and we must have $\|\boldsymbol{\lambda}_j\|_2 \leq 1$, $\|\boldsymbol{\lambda}_j\|_2 \leq 1$ for any j, k. From assumption 1, the right hand side of the previous inequality tends to zero as n tends to infinity. As such we conclude that $\left|\operatorname{cov}\left(h_1 m_{j1} \boldsymbol{\lambda}_j^\mathsf{T} \boldsymbol{u}_{(n)1}, h_1 m_{v1} \boldsymbol{\lambda}_v^\mathsf{T} \boldsymbol{u}_{(n)1}\right)\right|$ is $o(1)$. Some covariance terms appear in the expression for $\operatorname{var}(Z_{n1})$

$$\begin{aligned}\operatorname{var}(Z_{n1}) &= \frac{1}{k} \operatorname{var}\left(\sum_{j=1}^{k} h_1 m_{j1} \boldsymbol{\lambda}_j^\mathsf{T} \boldsymbol{u}_{(n)1}\right) \\ &= \frac{1}{k} \sum_{j=1}^{k} \operatorname{var}\left(h_i m_{ji} \boldsymbol{\lambda}_j^\mathsf{T} \boldsymbol{u}_{(n)1}\right) + \frac{1}{k} 2 \sum_{j=1}^{k-1} \sum_{v=j+1}^{k} \operatorname{cov}\left(h_1 m_{j1} \boldsymbol{\lambda}_j^\mathsf{T} \boldsymbol{u}_{(n)1}, h_1 m_{v1} \boldsymbol{\lambda}_v^\mathsf{T} \boldsymbol{u}_{(n)1}\right) \\ &= \frac{1}{k} \sum_{j=1}^{k} \operatorname{var}\left(h_1 m_{j1} \boldsymbol{\lambda}_j^\mathsf{T} \boldsymbol{u}_{(n)1}\right) + \frac{1}{k} 2 \sum_{j=1}^{k-1} \sum_{v=j+1}^{k} \boldsymbol{\lambda}_j^\mathsf{T} \boldsymbol{u}_{(n)1} \boldsymbol{\lambda}_v^\mathsf{T} \boldsymbol{u}_{(n)1} \\ &= \frac{1}{k} \sum_{j=1}^{k} \operatorname{var}\left(h_1 m_{j1} \boldsymbol{\lambda}_j^\mathsf{T} \boldsymbol{u}_{(n)1}\right) + o(1) \\ &= \frac{1}{k} \sum_{j=1}^{k} \boldsymbol{\lambda}_j^\mathsf{T} \boldsymbol{u}_{(n)1} \boldsymbol{u}_{(n)1}^\mathsf{T} \boldsymbol{\lambda}_j + o(1) \end{aligned} \tag{5.38}$$

The trailing term can be grouped into an $o(1)$ term as the sketch size k is fixed in our analysis. Using (5.37) and (5.38) we can then determine the row-wise variance totals s_n^2:

$$\begin{aligned}
s_n^2 &= \frac{1}{k}\sum_{i=1}^{r_n} \mathrm{var}(Z_{ni}) \\
&= \frac{1}{k}\sum_{i=1}^{r_n}\sum_{j=1}^{k} \boldsymbol{\lambda}_j^\mathsf{T}\boldsymbol{u}_{(n)i}\boldsymbol{u}_{(n)i}^\mathsf{T}\boldsymbol{\lambda}_j + o(1) \\
&= \frac{1}{k}\sum_{j=1}^{k} \boldsymbol{\lambda}_j^\mathsf{T}\left(\sum_{i=1}^{r_n}\boldsymbol{u}_{(n)i}\boldsymbol{u}_{(n)i}^\mathsf{T}\right)\boldsymbol{\lambda}_j + o(1) \\
&= \frac{1}{k}\sum_{j=1}^{k} \boldsymbol{\lambda}_j^\mathsf{T}\boldsymbol{U}_{(n)}^\mathsf{T}\boldsymbol{U}_{(n)}\boldsymbol{\lambda}_j + o(1) \\
&= \frac{1}{k}\sum_{j=1}^{k} \boldsymbol{\lambda}_j^\mathsf{T}\boldsymbol{I}_d\boldsymbol{\lambda}_j + o(1) \\
&= \frac{1}{k}\sum_{j=1}^{k} \boldsymbol{\lambda}_j^\mathsf{T}\boldsymbol{\lambda}_j + o(1) \\
&= \frac{1}{k} + o(1).
\end{aligned}$$

The step in the last line follows as we have taken $\boldsymbol{\lambda}$ to be a unit vector. The fact that $\boldsymbol{U}_{(n)}^\mathsf{T}\boldsymbol{U}_{(n)} = \boldsymbol{I}_d$ for all n serves as a useful normalisation to give stable limiting behaviour. We are working with a triangular array where the rows are nearly standardised. Asymptotically in n, $s_n^2 \to 1/k$.

We now establish a sequence of upper bounds (K_n). As the random variables in the construction of construction of the Hadamard sketch are bounded, we can bound the random variables in the triangular array (5.34) using the leverage scores of the sequence of source datasets. Now as the random sign $h_i \in \{+1, -1\}$ we have that for all $i = 1, \ldots, r_n$:

$$\begin{aligned}
|Z_{ni}| &= \frac{1}{\sqrt{k}}|h_i\sum_{j=1}^{k} m_{ji}\boldsymbol{\lambda}_j^\mathsf{T}\boldsymbol{u}_{(n)i}| \\
&= \frac{1}{\sqrt{k}}|\sum_{j=1}^{k} m_{ji}\boldsymbol{\lambda}_j^\mathsf{T}\boldsymbol{u}_{(n)i}|.
\end{aligned}$$

Now using the Cauchy-Schwarz inequality,

$$\frac{1}{\sqrt{k}}|\left(\sum_{j=1}^{k} m_{ji}\boldsymbol{\lambda}_j^\mathsf{T}\right)\boldsymbol{u}_{(n)i}| \leq \frac{1}{\sqrt{k}}\|\left(\sum_{j=1}^{k} m_{ji}\boldsymbol{\lambda}_j\right)\|_2\|\boldsymbol{u}_{(n)i}\|_2. \tag{5.39}$$

Using the triangle inequality,

$$\|\left(\sum_{j=1}^{k} m_{ji}\boldsymbol{\lambda}_j\right)\|_2 \leq \sum_{j=1}^{k}\|m_{ji}\boldsymbol{\lambda}_j\|_2. \tag{5.40}$$

Now as $m_{ji} \in \{+1, -1\}$ for all $j = 1, \ldots, k$,

$$\sum_{j=1}^{k}\|m_{ji}\boldsymbol{\lambda}_j\|_2 = \sum_{j=1}^{k}\|\boldsymbol{\lambda}_j\|_2. \tag{5.41}$$

As $\boldsymbol{\lambda}$ is a unit vector we can easily form the bound

$$\sum_{j=1}^{k}\|\boldsymbol{\lambda}_j\|_2 \leq k. \tag{5.42}$$

Substituting (5.41) and (5.42) into (5.39) leads to the upper bound for $i = 1, \ldots, r_n$:

$$\begin{aligned} |Z_{ni}| &\leq \frac{1}{\sqrt{k}} k \|\boldsymbol{u}_{(n)i}\|_2 \\ &= \sqrt{k} \|\boldsymbol{u}_{(n)i}\|_2. \end{aligned} \tag{5.43}$$

This is less than ideal as the upper bound is a function of k. In the present analysis the sketch size k is fixed. In future work we would like establish limit theorems letting k and d grow. This is discussed more in Chapter 6. We would like to eliminate the sketch size k from the upper bound (5.43). To do so we establish a tighter bound than in (5.42). It is reasonable to form a tighter bound as we have the restriction that $\boldsymbol{\lambda}$ is a unit vector, and we have not made full use of that in (5.42). To improve (5.42) we use the following lemma

Lemma 5.4. *Let $a_1, \ldots, a_k$ be a sequence of positive scalars such that $\sum_{j=1}^k a_j = 1$. Define $f(a_1, \ldots, a_k) = \sum_{j=1}^k a_j^{1/2}$. Then f s maximised by setting $a_j = 1/k$ for $j = 1, \ldots, k$. Furthermore $f(1/k, \ldots, 1/k) = \sqrt{k}$.*

The proof is simple using Lagrange multipliers. The quantity in (5.41) can be upper bounded using Lemma 5.4. We set $a_j = \|\boldsymbol{\lambda}_j\|_2^2$ for $j = 1, \ldots, k$. As $\boldsymbol{\lambda}$ is a unit vector we have that $\sum_{j=1}^k \|\boldsymbol{\lambda}_j\|_2^2 = \sum_{j=1}^k a_j = 1$. We can write $\sum_{j=1}^k \|\boldsymbol{\lambda}_j\|_2 = \sum_{j=1}^k a_j^{1/2}$. Take $f(a_1, \ldots, a_k) = \sum_{j=1}^k a_j^{1/2}$. We can form an upper bound on $\sum_{j=1}^k \|\boldsymbol{\lambda}_j\|_2$ by maximising f over the arguments $a_1, \ldots, a_k$ subject to the constraint that $\sum_{j=1}^k a_j = 1$. From Lemma 5.4 we have that

$$\sum_{j=1}^k \|\boldsymbol{\lambda}_j\|_2 \leq \sqrt{k}. \tag{5.44}$$

We can then form the tighter bounds on the triangular array

$$|Z_{ni}| \leq \frac{1}{\sqrt{k}} \sqrt{k} \|\boldsymbol{u}_{(n)i}\|_2 \tag{5.45}$$

$$= \|\boldsymbol{u}_{(n)i}\|_2. \tag{5.46}$$

The upper bound no longer depends on the sketch size k as was desired. Continuing, we can then form the sequence of upper bounds K_n:

$$K_n = \max_{i=1,\ldots,r_n} \|\boldsymbol{u}_{(n)i}\|_2.$$

We have that $|Z_{ni}| \leq K_n$ almost surely for $i = 1, \ldots, r_n$ and $n \in \mathbb{N}$. Assumption 1 (recall equation (5.21)) gives the limiting behaviour of K_n.

$$\begin{aligned} \lim_{n\to\infty} K_n &= \lim_{n\to\infty} \max_{i=1,\ldots,r_n} \|\boldsymbol{u}_{(n)i}\|_2 \\ &= 0. \end{aligned}$$

We have that $K_n \to 0$ as $n \to \infty$. As $s_n^2 = 1/k + o(1)$ we have an asymptotically standardised array and we can use Theorem 5.2 to conclude that the triangular array of random variables defined in (5.34) satisfies Lindeberg's condition. As such the conditions of Theorem 5.3 are satisfied. We conclude that the row sums in (5.35) converge in distribution to $N(0, 1/k)$. Finally, the Cramér-Wold device gives that the whitened sketched dataset has a limiting matrix normal distribution. That is the sequence of random matrices $\widetilde{\boldsymbol{A}}\boldsymbol{V}_{(n)}\boldsymbol{D}_{(n)}^{-1}$ converges in distribution to a $MN(\boldsymbol{0}, \boldsymbol{I}_k, \boldsymbol{I}_d/k)$ random matrix.

5.7 Proof of Theorem 4.5 (Complete sketching asymptotics)

Assumption 2:

$$\lim_{n\to\infty} n^{-1} \begin{bmatrix} \boldsymbol{y}_{(n)}^\top \boldsymbol{y}_{(n)} & \boldsymbol{y}_{(n)}^\top \boldsymbol{X}_{(n)} \\ \boldsymbol{X}_{(n)}^\top \boldsymbol{y}_{(n)} & \boldsymbol{X}_{(n)}^\top \boldsymbol{X}_{(n)} \end{bmatrix} = \boldsymbol{Q} \qquad \text{for some positive-definite matrix } \boldsymbol{Q}.$$

Theorem. *Suppose that Assumptions 1 and 2 hold, $k \geq p$, and $\boldsymbol{\beta}_S$ is computed using a Hadamard or Clarkson-Woodruff sketch. Let $(\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})^+$ denote the Moore-Penrose pseudo-inverse of $(\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})$. Let*

$$\widetilde{\boldsymbol{H}}_{(n)} = \frac{RSS_F^{(n)}}{k}\left(\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}}\right)^+ \text{ and } \boldsymbol{H}_{(n)} = \frac{RSS_F^{(n)}}{k-p+1}\left(\boldsymbol{X}_{(n)}^\mathsf{T}\boldsymbol{X}_{(n)}\right)^{-1}.$$

Then as $n \to \infty$, convergence in distribution holds for

$$(i)[\boldsymbol{H}_{(n)}^{-1/2}(\boldsymbol{\beta}_S - \boldsymbol{\beta}_F^{(n)})|\boldsymbol{A}_{(n)}] \to \text{Student}\left(\boldsymbol{0}, \boldsymbol{I}_p,\ k-p+1\right),$$
$$(ii)[\widetilde{\boldsymbol{H}}_{(n)}^{-1/2}(\boldsymbol{\beta}_S - \boldsymbol{\beta}_F^{(n)})\ |\boldsymbol{A}_{(n)}] \to N\left(\boldsymbol{0}, \boldsymbol{I}_p\right).$$

Notation is slightly heavier in the proof compared to the main text for the sake of clarity. Again we do not explicitly condition on the source dataset $\boldsymbol{A}_{(n)}$, the source dataset is always fixed, and the only randomness is from the sketching matrix. The sketched data will be denoted $\widetilde{\boldsymbol{y}}_{(n)}$ and $\widetilde{\boldsymbol{X}}_{(n)}$ to denote the dependence on the $n \times d$ source dataset. So $\widetilde{\boldsymbol{y}}_{(n)} = \boldsymbol{S}\boldsymbol{y}_{(n)}$ and $\widetilde{\boldsymbol{X}}_{(n)} = \boldsymbol{S}\boldsymbol{X}_{(n)}$. The dimension of the sketched dataset does not change.

Assumption 2 is of assistance in establishing the limit theorem. Let

$$\boldsymbol{Q}_{(n)} = n^{-1}\begin{bmatrix}\boldsymbol{y}_{(n)}^\mathsf{T}\boldsymbol{y}_{(n)} & \boldsymbol{y}_{(n)}^\mathsf{T}\boldsymbol{X}_{(n)} \\ \boldsymbol{X}_{(n)}^\mathsf{T}\boldsymbol{y}_{(n)} & \boldsymbol{X}_{(n)}^\mathsf{T}\boldsymbol{X}_{(n)}.\end{bmatrix}$$

The matrix $\boldsymbol{Q}_{(n)}$ contains the sufficient statistics needed to fit a Gaussian linear model, $\boldsymbol{y}_{(n)}^\mathsf{T}\boldsymbol{y}_{(n)}, \boldsymbol{X}_{(n)}^\mathsf{T}\boldsymbol{y}_{(n)}$ and $\boldsymbol{X}_{(n)}^\mathsf{T}\boldsymbol{X}_{(n)}$ given the source dataset $\boldsymbol{A}_{(n)} = [\boldsymbol{y}_{(n)}, \boldsymbol{X}_{(n)}]$. Assumption 2 states the averaged sufficient statistic matrix converges to a limiting matrix $\boldsymbol{Q}$. It will be helpful to partition the limiting matrix $\boldsymbol{Q}$ as

$$\boldsymbol{Q} = \lim_{n\to\infty} n^{-1}\begin{bmatrix}\boldsymbol{y}_{(n)}^\mathsf{T}\boldsymbol{y}_{(n)} & \boldsymbol{y}_{(n)}^\mathsf{T}\boldsymbol{X}_{(n)} \\ \boldsymbol{X}_{(n)}^\mathsf{T}\boldsymbol{y}_{(n)} & \boldsymbol{X}_{(n)}^\mathsf{T}\boldsymbol{X}_{(n)}\end{bmatrix} = \begin{bmatrix}s & \boldsymbol{m}^\mathsf{T} \\ \boldsymbol{m} & \boldsymbol{G}\end{bmatrix}, \tag{5.47}$$

where s is a scalar, $\boldsymbol{G}$ is a $p \times p$ matrix and $\boldsymbol{b}$ is a p-length column vector. The matrix $\boldsymbol{G}$ is the limiting averaged Gram matrix of the predictors. The vector $\boldsymbol{m}$ is the limit of the predictor response inner products $n^{-1}\boldsymbol{X}_{(n)}^\mathsf{T}\boldsymbol{y}_{(n)}$, and the scalar s is the limit of the mean total sum of squares $n^{-1}\boldsymbol{y}_{(n)}^\mathsf{T}\boldsymbol{y}_{(n)}$.

As mentioned, the assumption of a sequence of source datasets also gives a sequence of optimal least squares coefficients and residual errors. Let $\sigma_F^{2(n)} = RSS_F/n$. Define the limiting least squares coefficient estimate as $\boldsymbol{\beta} = \lim_{n\to\infty} \boldsymbol{\beta}_F^{(n)}$ and the limiting residual error as $\sigma^2 = \lim_{n\to\infty} \sigma_F^{2(n)}$. Both $\boldsymbol{\beta}$ and σ^2 can be expressed as functions of the matrix $\boldsymbol{Q}$. Specifically,

$$\boldsymbol{\beta} = \boldsymbol{G}^{-1}\boldsymbol{m}, \tag{5.48}$$
$$\sigma^2 = s - \boldsymbol{m}^\mathsf{T}\boldsymbol{G}^{-1}\boldsymbol{m}. \tag{5.49}$$

From Assumption 2, we have that $n^{-1}\boldsymbol{V}_{(n)}\boldsymbol{D}_{(n)}^2\boldsymbol{V}_{(n)}^\mathsf{T} \to \boldsymbol{Q}$. As such we have that $n^{-1/2}\boldsymbol{D}_{(n)}\boldsymbol{V}_{(n)}^\mathsf{T} \to \boldsymbol{Q}^{1/2}$ From the sketching central limit theorem the whitened sketched data converges to a matrix normal distribution

$$[\widetilde{\boldsymbol{y}}_{(n)}, \widetilde{\boldsymbol{X}}_{(n)}]\boldsymbol{V}_{(n)}\boldsymbol{D}_{(n)}^{-1} \xrightarrow{d} MN\left(\boldsymbol{0}_{k\times d}, \boldsymbol{I}_k, \boldsymbol{I}_d/k\right)$$

The benefit of adding Assumption 2 is that using Slutsky's theorem we have the additional convergence result

$$n^{-1/2}[\widetilde{\boldsymbol{y}}_{(n)}, \widetilde{\boldsymbol{X}}_{(n)}] \xrightarrow{d} MN\left(\boldsymbol{0}_{k\times d}, \boldsymbol{I}_k, \boldsymbol{Q}/k\right).$$

To prove results (i) and (ii) we use the continuous mapping theorem (Van Der Vaart, 1998, p. 7) in conjunction with the previous convergence result. It will be helpful to define the random variables $\widetilde{\boldsymbol{y}}_L, \widetilde{\boldsymbol{X}}_L$ as having the above limiting matrix normal distribution

$$[\widetilde{\boldsymbol{y}}_L, \widetilde{\boldsymbol{X}}_L] \sim MN\left(\boldsymbol{0}_{k\times d}, \boldsymbol{I}_k, \boldsymbol{Q}/k\right).$$

This is so we can say that

$$n^{-1/2}[\widetilde{\boldsymbol{y}}_{(n)}, \widetilde{\boldsymbol{X}}_{(n)}] \xrightarrow{d} [\widetilde{\boldsymbol{y}}_L, \widetilde{\boldsymbol{X}}_L].$$

Lemma 5.5 (Continuous Mapping Theorem)**.** *Let $\boldsymbol{Z}_n$ indicate a sequence of random vectors and $\boldsymbol{Z}$ indicate another random vector. Suppose the function $g : \mathbb{R}^d \to \mathbb{R}^m$ is continuous at every point of a set C such that $P(\boldsymbol{Z} \in C) = 1$. Then if $\boldsymbol{Z}_n \xrightarrow{d} \boldsymbol{Z}$ then $g(\boldsymbol{Z}_n) \xrightarrow{d} g(\boldsymbol{Z})$.*

In Lemma 5.5, the function $g : \mathbb{R}^d \to \mathbb{R}^m$ does not change with n, and the dimensions d and m are fixed when taking limits. The sketched estimator $\boldsymbol{\beta}_S$ can be defined as a function of the sketched data that is continuous over the set where $\widetilde{\boldsymbol{X}}_{(n)}$ is of full rank. Formally we could say that $\boldsymbol{\beta}_S = g(n^{-1/2}\widetilde{\boldsymbol{y}}_{(n)}, n^{-1/2}\widetilde{\boldsymbol{X}}_{(n)})$. As $\widetilde{\boldsymbol{X}}_L$ is of rank p almost surely, and $\widetilde{\boldsymbol{X}}_{(n)} \xrightarrow{d} \widetilde{\boldsymbol{X}}_L$ we can apply the continuous mapping theorem to determine the limiting distribution of the $\boldsymbol{\beta}_S$. The random matrix $[\widetilde{\boldsymbol{y}}_L, \widetilde{\boldsymbol{X}}_L]$ can be described using a hierarchical model completely analogous in structure to the hierarchical model established for the Gaussian sketch in section 3 of the main text. Specifically,

$$\widetilde{\boldsymbol{y}}_L \mid \widetilde{\boldsymbol{X}}_L \sim N\left(\widetilde{\boldsymbol{X}}\boldsymbol{\beta}, \frac{1}{k}\sigma^2 \boldsymbol{I}_k\right),$$
$$\widetilde{\boldsymbol{X}}_L \sim MN\left(\boldsymbol{0}_{k\times p}, \boldsymbol{I}_k, \frac{1}{k}\boldsymbol{Q}\right).$$

From Theorem 2 in the main text, and recalling that the function g outputs $\boldsymbol{\beta}_S$, we have that

$$g(\widetilde{\boldsymbol{y}}_L, \widetilde{\boldsymbol{X}}_L) \sim \text{Student}\left(\boldsymbol{\beta}, \frac{\sigma^2}{k-p+1}\boldsymbol{G}^{-1}, k-p+1\right).$$

As such, for the Hadamard and Clarkson-Woodruff sketches,

$$[\boldsymbol{\beta}_S \mid \boldsymbol{y}_{(n)}, \boldsymbol{X}_{(n)}] \xrightarrow{d} \text{Student}\left(\boldsymbol{\beta}, \frac{\sigma^2}{k-p+1}\boldsymbol{G}^{-1}, k-p+1\right).$$

Let

$$\boldsymbol{H}_{(n)} = \sigma_F^{2(n)}/(k-p+1)\left(n^{-1}\boldsymbol{X}_{(n)}^\mathsf{T}\boldsymbol{X}_{(n)}\right)^{-1}.$$

Now as $n^{-1}\boldsymbol{X}_{(n)}^\mathsf{T}\boldsymbol{X}_{(n)} \to \boldsymbol{G}$, $\sigma_F^{2(n)} \to \sigma^2$, and $\boldsymbol{\beta}_F^{(n)} \to \boldsymbol{\beta}$, Slutsky's theorem can be used to arrive at (i),

$$\boldsymbol{H}_{(n)}^{-1/2}(\boldsymbol{\beta}_S - \boldsymbol{\beta}_F) \xrightarrow{d} \text{Student}\left(\boldsymbol{0}, \boldsymbol{I}_p,\ k-p+1\right).$$

For result (ii), let us define the function

$$f(n^{-1/2}\widetilde{\boldsymbol{y}}_{(n)}, n^{-1/2}\widetilde{\boldsymbol{X}}_{(n)}) = \left[n\left(\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}}\right)^+\right]^{-1/2}\left(\widetilde{\boldsymbol{X}}^+\widetilde{\boldsymbol{y}} - \boldsymbol{\beta}\right)$$
$$= \left[n\left(\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}}\right)^+\right]^{-1/2}(\boldsymbol{\beta}_S - \boldsymbol{\beta}).$$

This function transforms the $\boldsymbol{\beta}_S$ so that the output is uncorrelated. This function is also continuous over the set where $\widetilde{\boldsymbol{X}}_{(n)}$ is of rank p. Again using the fact that $\widetilde{\boldsymbol{X}}_L$ has rank p almost surely, it follows from the continuous mapping theorem that $f(n^{-1/2}\widetilde{\boldsymbol{y}}_{(n)}, n^{-1/2}\widetilde{\boldsymbol{X}}_{(n)}) \xrightarrow{d} f(\widetilde{\boldsymbol{y}}_L, \widetilde{\boldsymbol{X}}_L)$. Result (ii) in Theorem 4.2 also applies to the hierarchical model for $\widetilde{\boldsymbol{y}}_L, \widetilde{\boldsymbol{X}}_L$, and gives the distribution of the transformed $\boldsymbol{\beta}_S$ under the Gaussian sketch. The distribution of $f(\widetilde{\boldsymbol{y}}_L, \widetilde{\boldsymbol{X}}_L)$ will be

$$f(\widetilde{\boldsymbol{y}}_L, \widetilde{\boldsymbol{X}}_L) \sim N\left(\boldsymbol{0}, \frac{\sigma^2}{k}\boldsymbol{I}_p\right).$$

As such, for the Clarkson-Woodruff and Hadamard sketches,

$$\left[n\left(\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}}\right)^+\right]^{-1/2}(\boldsymbol{\beta}_S - \boldsymbol{\beta}) \xrightarrow{d} \text{N}\left(\boldsymbol{0}, \frac{\sigma^2}{k}\boldsymbol{I}_p\right).$$

Now let

$$\widetilde{\boldsymbol{H}}_{(n)} = n\sigma_F^{2(n)}/k \left(\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}}\right)^+$$

Now as $\sigma_F^{2(n)} \to \sigma^2$, and $\boldsymbol{\beta}_F^{(n)} \to \boldsymbol{\beta}$, Slutsky's theorem can be used to arrive at (ii)

$$\widetilde{\boldsymbol{H}}_{(n)}^{-1/2}(\boldsymbol{\beta}_S - \boldsymbol{\beta}_F^{(n)}) \xrightarrow{d} N\left(\boldsymbol{0}, \boldsymbol{I}_p\right).$$

5.8 Proof of Theorem 4.6 (Partial sketching asymptotics)

Theorem. *Suppose that Assumptions 1, 2 and 3 hold, $k > p+3$, and $\boldsymbol{\beta}_P^*$ is computed using a Hadamard or Clarkson-Woodruff sketch. Let*

$$\boldsymbol{H}_{(n)} = \frac{(k-p-1)}{(k-p)(k-p-3)}\left(MSS_F^{(n)}(\boldsymbol{X}_{(n)}^\mathsf{T}\boldsymbol{X}_{(n)})^{-1} + \boldsymbol{\beta}_F^{(n)}\boldsymbol{\beta}_F^{(n)\mathsf{T}}\right).$$

Then as $n \to \infty$,

$$(i)\ E_S[\boldsymbol{\beta}_P^* - \boldsymbol{\beta}_F^{(n)}|\boldsymbol{A}_{(n)}] \to \boldsymbol{0}.$$
$$(ii)\ \mathrm{var}_S\left(\boldsymbol{H}_{(n)}^{-1/2}(\boldsymbol{\beta}_P^* - \boldsymbol{\beta}_F^{(n)})\ |\boldsymbol{A}_{(n)}\right) \to \boldsymbol{I}_d$$

Application of the continuous mapping theorem gives that the distribution of $\boldsymbol{\beta}_S$ and $\boldsymbol{\beta}_P^*$ under the Hadamard and Clarkson-Woodruff sketches converges to the distribution of the estimators under the Gaussian sketch. This does not necessarily guarantee convergence in moments. To establish a limit theorem for the bias and variance of the estimators, we need a uniform integrability condition on the sketched dataset. The sketched data will be denoted $\widetilde{\boldsymbol{X}}_{(n)}$ to denote the dependence on the $n \times p$ source covariate matrix. So $\widetilde{\boldsymbol{X}}_{(n)} = \boldsymbol{S}\boldsymbol{X}_{(n)}$. We again do not explicitly condition on the source dataset $\boldsymbol{A}_{(n)} = [\boldsymbol{y}_{(n)}, \boldsymbol{X}_{(n)}]$ in the following working.

Let $\boldsymbol{G}_{(n)} = n^{-1}\widetilde{\boldsymbol{X}}_{(n)}^\mathsf{T}\widetilde{\boldsymbol{X}}_{(n)}$. From the continuous mapping theorem and Theorem 4.3, it is known that

$$\boldsymbol{G}_{(n)}^{-1} \xrightarrow{d} \boldsymbol{W},$$

where $\boldsymbol{W}$ has an Inverse-Wishart$(k, k\boldsymbol{Q}^{-1})$ distribution and $\boldsymbol{Q}$ is the limiting matrix from assumption 2. We would like to establish convergence in first and second moments, that is

$$\mathbb{E}[\boldsymbol{G}_{(n)}^{-1}] \to \mathbb{E}[\boldsymbol{W}],$$
$$\mathrm{var}(\boldsymbol{G}_{(n)}^{-1}) \to \mathrm{var}(\boldsymbol{W}).$$

If convergence in first and second moments occurs, then we can show that (i) and (ii) will hold. If $\mathbb{E}[\boldsymbol{G}_{(n)}^{-1}] \to \mathbb{E}[\boldsymbol{W}]$, we can say that

$$\mathbb{E}[\boldsymbol{\beta}_P^* - \boldsymbol{\beta}] \to \boldsymbol{0},$$

where $\boldsymbol{\beta}$ is the limiting ordinary least squares estimator (5.48), that is a function of the limiting matrix $\boldsymbol{Q}$ in assumption 2. From here, using that $\lim_{n\to\infty} \boldsymbol{\beta}_F^{(n)}$, Slutsky's theorem can be used to arrive at (i)

$$\mathbb{E}[\boldsymbol{\beta}_P^* - \boldsymbol{\beta}_F^{(n)}] \to \boldsymbol{0}.$$

To show convergence of the variance of the sketched estimator (ii), we define

$$\boldsymbol{H}_{(n)} = \frac{(k-p-1)}{(k-p)(k-p-3)}\left(MSS_F^{(n)}\left(\boldsymbol{X}_{(n)}^\mathsf{T}\boldsymbol{X}_{(n)}\right)^{-1} + \frac{(k-p+1)}{(k-p-1)}\boldsymbol{\beta}_F^{(n)}\boldsymbol{\beta}_F^{(n)\mathsf{T}}\right),$$
$$\boldsymbol{H} = \frac{(k-p-1)}{(k-p)(k-p-3)}\left((s-\sigma^2)\boldsymbol{G}^{-1} + \frac{(k-p+1)}{(k-p-1)}\boldsymbol{\beta}\boldsymbol{\beta}^\mathsf{T}\right).$$

Where s, σ^2 and and $\boldsymbol{G}$ are functions of the limiting matrix $\boldsymbol{Q}$, as in (5.47), (5.48) and (5.49). If $\text{var}(\boldsymbol{G}_{(n)}^{-1}) \to \text{var}(\boldsymbol{W})$ it follows that

$$\text{var}_S\left(\boldsymbol{H}^{-1/2}(\boldsymbol{\beta}_P^* - \boldsymbol{\beta})\right) \to \boldsymbol{I}_d.$$

As $\boldsymbol{H}_{(n)}$ converges to $\boldsymbol{H}$ and $\boldsymbol{\beta}_F^{(n)}$ converges to $\boldsymbol{\beta}$ asymptotically with n, an application of Slutsky's theorem gives (ii),

$$\text{var}_S\left(\boldsymbol{H}_{(n)}^{-1/2}(\boldsymbol{\beta}_P^* - \boldsymbol{\beta}_F^{(n)})\right) \to \boldsymbol{I}_d$$

As such, if we can establish that $\text{var}(\boldsymbol{G}_{(n)}^{-1}) \to \text{var}(\boldsymbol{W})$ we have proved (ii). The following theorem describes the necessary conditions for such convergence to occur.

Theorem 5.4. *(Billinglsley, 1968, Theorem 5.4) Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a sequence of random vectors. Suppose $\boldsymbol{X}_n$ converges in distribution to a random variable $\boldsymbol{Z}$ as n tends to infinity. For the additional convergence of moments $\mathbb{E}[\boldsymbol{X}_n] \to \mathbb{E}[\boldsymbol{Z}]$ and var$[\boldsymbol{X}_n] \to$ var$[\boldsymbol{Z}]$, it must hold that for all conformable constant vectors $\boldsymbol{\lambda}$*

$$\lim_{M\to\infty} \limsup_{n\to\infty} |\boldsymbol{\lambda}^\intercal \boldsymbol{X}_n|^2 \mathbb{1}_{\{|\boldsymbol{\lambda}^\intercal \boldsymbol{X}_n|^2 \geq M\}} = 0.$$

The above condition can be difficult to verify directly. It can be shown that if asymptotically $|\boldsymbol{\lambda}^\intercal \boldsymbol{X}_n|$ has a bounded fourth moment, then the integrability condition is satisfied (Van Der Vaart, 1998, section 2.5).

A linear combination of the elements of the random matrix $\boldsymbol{G}_{(n)}^{-1}$ can be written an trace$(\boldsymbol{\Lambda}\boldsymbol{G}_{(n)}^{-1})$ for a $p \times p$ matrix of constants $\boldsymbol{\Lambda}$. It is easier to work with this form rather than stacking the elements of the random matrix to form a random vector. From theorem 5.4, it is sufficient to show that show that the expected value of $|\text{trace}(\boldsymbol{\Lambda}\boldsymbol{G}_{(n)}^{-1})|^4$ is finite for large n to show the desired convergence in moments.

As trace$(\boldsymbol{\Lambda}\boldsymbol{G}_{(n)}^{-1})$ equals the sum of the singular values of the matrix $\boldsymbol{\Lambda}\boldsymbol{G}_{(n)}^{-1}$, we can form an upper bound on the value,

$$\begin{aligned}\text{trace}(\boldsymbol{\Lambda}\boldsymbol{G}_{(n)}^{-1}) &\leq p\|\boldsymbol{\Lambda}\boldsymbol{G}_{(n)}^{-1}\|_2 \\ &\leq p\|\boldsymbol{\Lambda}\|_2\|\boldsymbol{G}_{(n)}^{-1}\|_2\end{aligned}$$

Squaring both sides gives an upper bound on the quantity that must satisfy the uniform integrability condition,

$$|\text{trace}(\boldsymbol{\Lambda}\boldsymbol{G}_{(n)}^{-1})|^2 \leq p^2\|\boldsymbol{\Lambda}\|_2^2\|\boldsymbol{G}_{(n)}^{-1}\|_2^2.$$

Squaring again gives an upper bound on the fourth moment of the linear combination of interest

$$\begin{aligned}|\text{trace}(\boldsymbol{\Lambda}\boldsymbol{G}_{(n)}^{-1})|^4 &\leq p^4\|\boldsymbol{\Lambda}\|_2^4\|\boldsymbol{G}_{(n)}^{-1}\|_2^4 \\ &= p^4\|\boldsymbol{\Lambda}\|_2^4\left(\frac{1}{\sigma^2_{\min}(\boldsymbol{G}_{(n)})}\right)^2\end{aligned}$$

By assumption 3, the expectation of the right hand side is finite. As such, the uniform integrability condition holds and we can conclude that

$$\begin{aligned}\mathbb{E}[\boldsymbol{G}_{(n)}^{-1}] &\to \mathbb{E}[\boldsymbol{W}], \\ \text{var}(\boldsymbol{G}_{(n)}^{-1}) &\to \text{var}(\boldsymbol{W}).\end{aligned}$$

As discussed at the beginning of the proof this is sufficient to show that (i) and (ii) hold.

6

On subspace embeddings, Tracy-Widom limits and approximate Bayesian subset selection

6.1 Summary

Sketching is a probabilistic data compression technique that uses random projection. Sketching has recently been proposed as a method for approximate Bayesian regression on large datasets. We investigate sketching for approximate Bayesian subset selection. Sketching algorithms offer probabilistic bounds on the discrepancy introduced from using the sketched dataset in place of the full dataset. Existing bounds give limited information on how to choose the size of the compressed dataset. We consider the asymptotic behaviour of various random sketching matrices and establish an important link between the Tracy-Widom distribution and the stochastic error rates of sketching algorithms. The Tracy-Widom distribution can be used to construct asymptotic error bounds describing the information loss from the use of the randomised algorithm. The asymptotic results provide new insights on the comparative performance of different sketching algorithms and will help to give useful guidelines for practitioners. We test the theory and methods on a large genetic dataset.

6.2 Introduction

As discussed in Chapter 4 sketching algorithms use random projection to generate a smaller surrogate dataset for efficient approximate computation. Recent work has examined the possibility of using sketching for approximate Bayesian inference (Geppert et al., 2017; Bardenet and Maillard, 2015). An appreciable barrier to the adoption of inferential procedures using sketching is the difficulty in constructing tight error bounds on the sketching noise injected into the analysis. We take a new approach and use random matrix asymptotics to assess the suitability of random projections for approximate Bayesian inference. A main finding is that under mild regularity conditions, the stochastic distortion caused by a wide class of random projections is well described by the Tracy-Widom law (Tracy and Widom, 1994). The asymptotic results help to determine appropriate compression ratios for sketched analyses and to compare the operating characteristics of different random projections. We assess the suitability of sketching for approximate Bayesian subset selection using a combination of asymptotic theory and simulation.

We assume we have n independent observations from the standard Gaussian linear model. For simplicity the error variance σ^2 is treated as known. Let $\boldsymbol{y}$ denote a vector of n responses, and let $\boldsymbol{X}$ denote a $n \times p$ matrix of covariates. As discussed in Chapter 4, sketching algorithms use random linear mappings to the size of the dataset from n to k observations. The random linear mapping can be represented as a $k \times n$ sketching matrix $\boldsymbol{S}$. Complete sketching generates a k-length sketched response vector $\widetilde{\boldsymbol{y}}$ and a $k \times p$ matrix of sketched predictors $\widetilde{\boldsymbol{X}}$. The sketched data are computed through the linear mappings $\widetilde{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$ and $\widetilde{\boldsymbol{X}} = \boldsymbol{S}\boldsymbol{X}$. Partial sketching only generates a $k \times p$ matrix of sketched covariates $\widetilde{\boldsymbol{X}}$. We again use the random mapping $\widetilde{\boldsymbol{X}} = \boldsymbol{S}\boldsymbol{X}$. To simplify the analysis we only consider complete sketching in this chapter. Extensions to partial sketching are covered in the discussion. Working with the sketched

responses $\widetilde{\boldsymbol{y}}$ and sketched covariates $\widetilde{\boldsymbol{X}}$ is more computationally efficient than operating on the full dataset of n observations. The use of structured random projections $\boldsymbol{S}$ allows one to quantify the information loss incurred from working with the compressed dataset.

As discussed in section 4.2.1 of Chapter 4, sketching algorithms are typically motivated from objects known as ϵ-subspace embeddings (Woodruff, 2014; Meng and Mahoney, 2013; Yang et al., 2015a). Recall the formal definition of an ϵ-subspace embedding.

Definition 6.1. ϵ-subspace embedding.
For a given $n \times p$ matrix $\boldsymbol{A}$, we call a $k \times n$ matrix $\boldsymbol{S}$ an ϵ-subspace embedding for $\boldsymbol{A}$, if for all vectors $\boldsymbol{z} \in \mathbb{R}^p$

$$(1-\epsilon)||\boldsymbol{A}\boldsymbol{z}||_2^2 \leq ||\boldsymbol{S}\boldsymbol{A}\boldsymbol{z}||_2^2 \leq (1+\epsilon)||\boldsymbol{A}\boldsymbol{z}||_2^2.$$

An ϵ-subspace preserves the linear structure of the original dataset up to a multiplicative $(1 \pm \epsilon)$ factor. Broadly speaking, the covariance matrix of the sketched dataset is similar to the covariance matrix of the source dataset if ϵ is small. Mathematical arguments show that the sketched dataset is a good surrogate for many linear statistical methods if the sketching matrix $\boldsymbol{S}$ is an ϵ-subspace embedding for the original dataset, with ϵ sufficiently small (Woodruff, 2014). Epsilon-subspace embeddings are central in our analysis for developing posterior approximations and for constructing probabilistic bounds on the compression loss. Suitable ranges for ϵ depend on the task of interest and structural properties of the source dataset.

Sketching has been investigated for posterior approximation in the fixed model setting. The likelihood is $p(\boldsymbol{y}|\boldsymbol{X}, \beta, \sigma^2) = N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$, where $\boldsymbol{\beta}$ is a p-dimensional vector of coefficients. Given a prior distribution $p(\boldsymbol{\beta})$, the target posterior distribution of interest is $p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^2) \propto p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta})$. For large n, it may be computationally expensive to sample from the posterior distribution using Markov-Chain Monte Carlo, or to determine the analytic form if a conjugate prior is used. Geppert et al. (2017) consider using the sketched dataset $[\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}]$ to construct an approximate posterior distribution on the coefficients $\boldsymbol{\beta}$. Given an ϵ-subspace embedding it is possible to establish bounds on the difference between the sketched posterior and the target posterior as a function of ϵ (Geppert et al., 2017). We consider approximate Bayesian model selection, also motivated by properties of ϵ-subspace embeddings.

To assess the confidence in the approximate posterior it is necessary to obtain the probability that the random projection $\boldsymbol{S}$ is an ϵ-subspace embedding for the source dataset. The embedding probability

$$\Pr(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}), \tag{6.1}$$

is a critical feature of many sketching algorithms that is difficult to characterise precisely using existing theory (Venkatasubramanian and Wang, 2011). Most existing results on the embedding probability are finite sample lower bounds, that can potentially contain hidden constants. This makes it difficult to design methods for reporting the level of uncertainty associated with the sketched approximation. We take a new approach and develop asymptotic expressions for the embedding probability (6.1). We find that the Tracy-Widom law can be used to describe the probability of obtaining an ϵ-subspace embedding for many data oblivious sketches in the literature. The Tracy-Widom law has many applications in high-dimensional statistics (Johnstone, 2006; Bai and Silverstein, 2010), however there is little work to our knowledge foregrounding the importance of the law for statistical computation. The connection to sketching algorithms is novel and hints at deeper principles that may apply to randomised algorithms that use data oblivious projections.

We test the accuracy of the asymptotic results on a large genetic dataset and find that the Tracy-Widom law gives a very good approximation for the embedding probability (6.1). We also analyse a real genetic dataset to provide guidelines on the necessary ϵ required for approximate Bayesian model selection. It appears that the noise introduced by the sketch can overwhelm the posterior distribution in situations where there is model uncertainty. False positives are also another systemic issue with sketched model

selection. Although single pass sketching does not appear to be a viable route for posterior approximation, we believe the asymptotic analysis presented here will nevertheless be useful in the design and analysis of general sketching algorithms.

6.3 Bayesian model selection

We assume that the model is unknown, and wish to incorporate model uncertainty into the Bayesian analysis. The p-length binary vector $\boldsymbol{\gamma}$ is used to index different models. Element j in $\boldsymbol{\gamma}$ is equal to one if variable j is included in the model, and zero otherwise. Let $\boldsymbol{X}_\gamma$ denote the design matrix for a particular model. The matrix $\boldsymbol{X}_\gamma$ is a submatrix of $\boldsymbol{X}$ where column j is included if $\boldsymbol{\gamma}_j = 1$. Let p_γ denote the number of predictors in model $\boldsymbol{\gamma}$, p_γ is equal to the sum of the elements of $\boldsymbol{\gamma}$. The likelihood is defined as

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}_\gamma, \sigma^2, \boldsymbol{\gamma}) = N(\boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma, \sigma^2 \boldsymbol{I}_n),$$

where $\boldsymbol{\beta}_\gamma$ is a p_γ-dimensional vector of coefficients for model $\boldsymbol{\gamma}$. We will use Zellner's g-prior,

$$p(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}, \sigma^2) = N(\boldsymbol{0}, g\sigma^2(\boldsymbol{X}_\gamma^\mathsf{T}\boldsymbol{X}_\gamma)^{-1}),$$

where g is a hyper-parameter controlling the prior variance over the coefficients. As stated in the introduction we treat σ^2 as known. Let the residual sum of squares for a particular model $\boldsymbol{\gamma}$ be denoted

$$RSS_F^\gamma = \boldsymbol{y}^\mathsf{T}\boldsymbol{y} - \boldsymbol{y}^\mathsf{T}\boldsymbol{X}_\gamma(\boldsymbol{X}_\gamma^\mathsf{T}\boldsymbol{X}_\gamma)^{-1}\boldsymbol{X}_\gamma^\mathsf{T}\boldsymbol{y}.$$

The marginal likelihood is a function of the residual sum of squares RSS_F^γ and the total sum of squares $\boldsymbol{y}^\mathsf{T}\boldsymbol{y}$. The expression is

$$p(\boldsymbol{y}|\boldsymbol{\gamma}, g, \sigma^2) = (1+g)^{-p_\gamma/2} \exp\left[-\frac{1}{2\sigma^2}\left(\frac{g}{g+1}RSS_F^\gamma + \frac{1}{g+1}\boldsymbol{y}^\mathsf{T}\boldsymbol{y}\right)\right]. \tag{6.2}$$

We assume we use Markov chain Monte Carlo (MCMC) to sample from the posterior distribution over models, for example the evolutionary stochastic search algorithm in Bottolo and Richardson (2010). For tall datasets it is likely to be advantageous to precompute the sufficient statistics $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}, \boldsymbol{X}^\mathsf{T}\boldsymbol{y}, \boldsymbol{y}^\mathsf{T}\boldsymbol{y}$. Computing $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ is $O(np^2)$, computing $\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$ is $O(np)$ and computing $\boldsymbol{y}^\mathsf{T}\boldsymbol{y}$ is $O(n)$. This is a one-off set up cost of $O(np^2)$ operations, dominated by computation of the Gram matrix of predictors $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$.

At each step in the MCMC algorithm it is only necessary to compute the model sum of squares $\boldsymbol{y}^\mathsf{T}\boldsymbol{X}_\gamma(\boldsymbol{X}_\gamma^\mathsf{T}\boldsymbol{X}_\gamma)^{-1}\boldsymbol{X}_\gamma^\mathsf{T}\boldsymbol{y}$ for each new model $\boldsymbol{\gamma}$ visited. This can be obtained in $O(p_\gamma^3)$ operations. The first step is to compute $(\boldsymbol{X}_\gamma^\mathsf{T}\boldsymbol{X}_\gamma)^{-1}\boldsymbol{X}_\gamma^\mathsf{T}\boldsymbol{y}$ by solving the linear system

$$\boldsymbol{X}_\gamma^\mathsf{T}\boldsymbol{X}_\gamma\boldsymbol{b} = \boldsymbol{X}_\gamma^\mathsf{T}\boldsymbol{y} \tag{6.3}$$

for $\boldsymbol{b}$. The matrix $\boldsymbol{X}_\gamma^\mathsf{T}\boldsymbol{X}_\gamma$ is obtained by taking the appropriate subset from the full Gram matrix of the predictors $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$. Likewise, the marginal associations $\boldsymbol{X}_\gamma^\mathsf{T}\boldsymbol{y}$ are read from the summary statistic $\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$. Solving the linear system (6.3) takes $O(p_\gamma^3)$ operations.

Given the solution $\widehat{\boldsymbol{b}} = (\boldsymbol{X}_\gamma^\mathsf{T}\boldsymbol{X}_\gamma)^{-1}\boldsymbol{X}_\gamma^\mathsf{T}\boldsymbol{y}$, computation of the product $\boldsymbol{y}^\mathsf{T}\boldsymbol{X}_\gamma[(\boldsymbol{X}_\gamma^\mathsf{T}\boldsymbol{X}_\gamma)^{-1}\boldsymbol{X}_\gamma^\mathsf{T}\boldsymbol{y}]$ can be done in $O(p_\gamma)$ time, leading to an overall $O(p_\gamma^3)$ cost for computing the integrated likelihood. The pay-off from the initial investment in computing the sufficient statistics is that the subsequent cost of each MCMC step is then independent of n.

For a very tall dataset the $O(np^2)$ set up cost of computing $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ may be undesirable. To reduce the initial computational expense, we can approximate the sufficient statistics using the sketched dataset. Given a sketched dataset of k observations $[\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}]$, the sketched sufficient statistics $\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{y}}^\mathsf{T}\widetilde{\boldsymbol{y}}$ can be computed in $O(kp^2)$ time. The approximate posterior distribution over models is defined using the sketched sufficient statistics. The approximate posterior can be explored using the same MCMC algorithm that would be applied to the full dataset sufficient statistics. Use of the sketched sufficient statistics is formally motivated using ϵ-subspace embeddings in the next section.

6.4 Approximate Bayesian inference

Geppert et al. (2017) examine sketching for approximate Bayesian regression in the fixed model setting, using properties of ϵ-subspace embeddings to motivate their approach. If $\widetilde{\boldsymbol{A}} = [\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}]$ is an ϵ-subspace embedding of $\boldsymbol{A} = [\boldsymbol{y}, \boldsymbol{X}]$, it must hold that for all $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$(1-\epsilon)||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \leq ||\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}||_2^2 \leq (1+\epsilon)||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2. \tag{6.4}$$

Let $p(\boldsymbol{\beta})$ denote the prior distribution and $p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}, \sigma^2)$ denote the likelihood, so

$$p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2\right).$$

The target posterior distribution is $p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) \propto p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}, \sigma^2)p(\boldsymbol{\beta})$. If n is very large it may be computationally infeasible to compute or sample from the target posterior distribution. Geppert et al. propose to form an approximate likelihood function $\widetilde{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}, \sigma^2)$ by substituting the sketched squared residuals $||\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}||_2^2$ for the true squared residuals $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$,

$$\widetilde{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}||\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}||_2^2\right).$$

If ϵ is small, the bounds (6.4) imply that $\widetilde{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}, \sigma^2) \approx p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}, \sigma^2)$. Using the sketched likelihood we can then define a sketched posterior

$$\widetilde{p}(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y}, \sigma^2) \propto p(\boldsymbol{\beta})\widetilde{p}(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2).$$

As $\epsilon \to 0$, the approximate likelihood $\widetilde{p}(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)$ approaches the true likelihood $p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}, \sigma^2)$, and the sketched approximate posterior approaches the target posterior distribution. More formally, as $\epsilon \to 0$,

$$D(p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^2) \parallel \widetilde{p}(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^2)) \to 0.$$

where $D(p(\boldsymbol{\beta}|\boldsymbol{\beta}, \boldsymbol{X}, \sigma^2) \parallel \widetilde{p}(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^2))$ denotes the Kullback-Leibler divergence of the sketched posterior from the target posterior. Geppert et al. establish a bound on the Wasserstein distance between the sketched posterior and the target posterior for finite ϵ.

We can use a similar argument to construct an approximate posterior distribution over models. Let RSS_S^γ denote the residual sum of squares using the sketched dataset for a particular model $\boldsymbol{\gamma}$,

$$\begin{aligned} RSS_S^\gamma &= \min_{\boldsymbol{\beta} \in \mathbb{R}^{p_\gamma}} \|\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}_\gamma \boldsymbol{\beta}\|_2^2 \\ &= \widetilde{\boldsymbol{y}}^\mathsf{T}\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{y}}^\mathsf{T}\widetilde{\boldsymbol{X}}_\gamma(\widetilde{\boldsymbol{X}}_\gamma^\mathsf{T}\widetilde{\boldsymbol{X}}_\gamma)^{-1}\widetilde{\boldsymbol{X}}_\gamma^\mathsf{T}\widetilde{\boldsymbol{y}}. \end{aligned}$$

Lemma 6.1 gives an important result that motivates our approach.

Lemma 6.1. *Suppose $\widetilde{\boldsymbol{A}} = [\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}]$ is an ϵ-subspace embedding of $\boldsymbol{A} = [\boldsymbol{y}, \boldsymbol{X}]$. Then for all models γ*

$$(1-\epsilon)RSS_F^\gamma \leq RSS_S^\gamma \leq (1+\epsilon)RSS_F^\gamma.$$

Proof: Let $\boldsymbol{\beta}_F^\gamma$ denote the least squares coefficients for model $\boldsymbol{\gamma}$ using the full dataset, where omitted covariates have their respective coefficients set to zero. The vector $\boldsymbol{\beta}_F^\gamma$ is p dimensional, with p_γ nonzero elements. Similarly, let $\boldsymbol{\beta}_S^\gamma$ denote the least squares coefficients for model $\boldsymbol{\gamma}$ using the sketched dataset, where again omitted covariates have their respective coefficients set to zero. The vector $\boldsymbol{\beta}_S^\gamma$ is also p dimensional, with p_γ nonzero elements. It holds that $RSS_F^\gamma = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_F^\gamma\|_2^2$ and $RSS_S^\gamma = \|\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}_S^\gamma\|_2^2$. To establish the upper bound in Lemma 6.1 we use the upper bound in (6.4),

$$\begin{aligned} \|\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}_S^\gamma||_2^2 &\leq \|\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}_F^\gamma||_2^2 \\ &\leq (1+\epsilon)||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_F^\gamma||_2^2 \\ &= (1+\epsilon)RSS_F^\gamma. \end{aligned}$$

The first line follows from the optimality of $\boldsymbol{\beta}_S^\gamma$ on the sketched dataset. To establish the lower bound in Lemma 6.1 we use the lower bound in (6.4),

$$\begin{aligned}||\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}_S^\gamma||_2^2 &\geq (1-\epsilon)\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_S^\gamma||_2^2 \\ &\geq (1-\epsilon)\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_F^\gamma||_2^2 \\ &= (1-\epsilon)RSS_F^\gamma.\end{aligned}$$

The second line follows from the optimality of $\boldsymbol{\beta}_F^\gamma$ on the full dataset, completing the proof.

We can define an approximate integrated likelihood $\widetilde{p}(\boldsymbol{y}|\boldsymbol{\gamma}, g, \sigma^2)$ in terms of the residual sum of squares on the sketched dataset. We substitute RSS_S^γ in place of RSS_F^γ in the target integrated likelihood (6.2), obtaining

$$\widetilde{p}(\boldsymbol{y}|\boldsymbol{\gamma}, g, \sigma^2) = (1+g)^{-p_\gamma/2} \exp\left[-\frac{1}{2\sigma^2}\left(\frac{g}{g+1}RSS_S^\gamma + \frac{1}{g+1}\boldsymbol{y}^\mathsf{T}\boldsymbol{y}\right)\right]. \tag{6.5}$$

If $[\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}]$ is an ϵ-subspace embedding of $[\boldsymbol{y}, \boldsymbol{X}]$, the bounds on the residual sum of squares $(1-\epsilon)RSS_F^\gamma \leq RSS_S^\gamma \leq (1+\epsilon)RSS_F^\gamma$ must hold. If ϵ is small, the approximate integrated likelihood $\widetilde{p}(\boldsymbol{y}|\boldsymbol{\gamma}, g, \sigma^2)$ should be a good approximation of the target integrated likelihood $p(\boldsymbol{y}|\boldsymbol{\gamma}, g, \sigma^2)$. The sketched posterior approximation over models is then defined as

$$\widetilde{p}(\boldsymbol{\gamma}|\boldsymbol{y}, g, \sigma^2) \propto p(\boldsymbol{\gamma})\widetilde{p}(\boldsymbol{y}|\boldsymbol{\gamma}, g, \sigma^2).$$

As $\epsilon \to 0$, $\widetilde{p}(\boldsymbol{y}|\boldsymbol{\gamma}, g, \sigma^2)$ approaches $p(\boldsymbol{y}|\boldsymbol{\gamma}, g, \sigma^2)$, and the sketched approximate posterior approaches the target posterior distribution. More formally, as $\epsilon \to 0$,

$$D(p(\boldsymbol{\gamma}|\boldsymbol{y}, g, \sigma^2) \parallel \widetilde{p}(\boldsymbol{\gamma}|\boldsymbol{y}, g, \sigma^2)) \to 0,$$

where $D(p(\boldsymbol{\gamma}|\boldsymbol{y}, g, \sigma^2) \parallel \widetilde{p}(\boldsymbol{\gamma}|\boldsymbol{y}, g, \sigma^2))$ denotes the Kullback-Leibler divergence of the sketched posterior from the target posterior.

A Markov chain Monte Carlo algorithm targeting the sketched posterior requires a set up cost of $O(kp^2)$ operations, after which each MCMC step to a new model has $O(p_\gamma^3)$ cost. If we can reliably generate ϵ-subspace embeddings with small ϵ we can arrive at a posterior approximation with a smaller overall computational cost than through exact computation.

To form concrete probabilistic error bounds we need to determine the probability that the random sketching matrix $\boldsymbol{S}$ is an ϵ-subspace embedding for the source dataset $[\boldsymbol{y}, \boldsymbol{X}]$. The success probability (6.1) determines the confidence we have in the randomised algorithm for approximate regression. We develop useful guidelines for practitioners in section 4.

Secondly, we also need to know what value of ϵ is needed to obtain a tolerable posterior approximation. Geppert et al. (2017) aim for $\epsilon \in (0.1, 0.2)$ for posterior approximation in the fixed model setting. We examine what is an appropriate ϵ for posterior approximation through simulation. Our results suggest that ϵ needs to be much smaller than 0.1 for approximating the posterior distribution over models. We now turn to the first question of interest, determining the embedding probability.

6.5 Embedding probabilities

6.5.1 Previous work

We first review some key theoretical results used to support sketching algorithms from the computer science literature. Let $\boldsymbol{A}$ denote an arbitrary $n \times d$ data matrix that we wish to compress. Data oblivious random projections are distributions over sketching matrices $\boldsymbol{S} \in \mathbb{R}^{k \times n}$ that are not a function of the source dataset $\boldsymbol{A} \in \mathbb{R}^{n \times d}$. At a high-level, data oblivious random projections offer guarantees on the success probability

$$\Pr(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}),$$

Sketch	Sketching time	Required sketch size k
Gaussian	$O(ndk)$	$O((d+\log(1/\delta))/\epsilon^2)$
Hadamard	$O(nd\log k)$	$O((\sqrt{d}+\sqrt{\log n})^2(\log(d/\delta))/\epsilon^2)$
Clarkson-Woodruff	$O(nd)$	$O(d^2/\delta\epsilon^2)$

Table 6.1: Properties of different data oblivious random projections (Woodruff, 2014). The third column refers to the necessary sketch size k to obtain an ϵ-subspace embedding for an arbitrary $n \times d$ source dataset with at least probability $(1-\delta)$.

for an arbitrary $n \times d$ source dataset $\boldsymbol{A}$. The guarantees are typically finite sample lower bounds on the embedding probability. The bounds concern the required sketch size k needed to obtain an ϵ-subspace embedding with probability at least $1-\delta$. The lower bounds for the Gaussian, Hadamard and Clarkson-Woodruff sketches are listed in Table 6.1.

The results are expressed in Big-O notation, the formulae for the required sketch size k are not fully explicit. For the Gaussian sketch, the notation $O((\sqrt{d}+\sqrt{\log n})^2(\log(d/\delta))/\epsilon^2)$ indicates the existence of constants c_1 and c_2 that are independent of n, such that if k is chosen within the bounds

$$c_1(d+\log(1/\delta))/\epsilon^2) \leq k \leq c_2(d+\log(1/\delta))/\epsilon^2),$$

then the probability of obtaining an ϵ-subspace embedding is at least $(1-\delta)$. The hidden constants in Table 6.1 are of little concern from a computer science perspective as the algorithmic significance is that the required sketch size k is independent of n for the Gaussian and Clarkson-Woodruff sketches, and very weakly dependent on n for the Hadamard projection. This implies that size of the sketch does not necessarily have to increase with n to obtain a fixed error guarantee. This is a very desirable property for a data compression algorithm.

Many proofs of the results in Table 6.1 use the following important lemma that gives a necessary and sufficient condition to obtain an ϵ-subspace embedding (Woodruff, 2014, Chapter 2).

Lemma 6.2. *For a given matrix $n \times d$ matrix $\boldsymbol{A}$ of rank d, let $\boldsymbol{U}$ be an orthonormal basis for the columns of $\boldsymbol{A}$. Let $\boldsymbol{S}$ be some $k \times n$ sketching matrix. The matrix $\boldsymbol{S}$ an ϵ-subspace embedding for $\boldsymbol{A}$, if and only if*

$$\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}) \leq \epsilon,$$

where $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})$ denotes the maximum singular value of the $d \times d$ matrix $\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}$.

Proof: From Definition 6.1, it is necessary to preserve norms in the column space of $\boldsymbol{A}$ up to $(1 \pm \epsilon)$ factor,

$$(1-\epsilon)||\boldsymbol{A}\boldsymbol{z}||_2^2 \leq ||\boldsymbol{S}\boldsymbol{A}\boldsymbol{z}||_2^2 \leq (1+\epsilon)||\boldsymbol{A}\boldsymbol{z}||_2^2.$$

It is therefore sufficient to consider an orthonormal basis for the column space of $\boldsymbol{A}$, as

$$\left\{\boldsymbol{A}\boldsymbol{z} : \boldsymbol{z} \in \mathbb{R}^d\right\} = \left\{\boldsymbol{U}\boldsymbol{v} : \boldsymbol{v} \in \mathbb{R}^d\right\}.$$

We can limit attention to unit vectors $\boldsymbol{v}$, as if

$$(1-\epsilon)||\boldsymbol{U}\boldsymbol{v}||_2^2 \leq ||\boldsymbol{S}\boldsymbol{U}\boldsymbol{v}||_2^2 \leq (1+\epsilon)||\boldsymbol{U}\boldsymbol{v}||_2^2,$$

holds for all unit vectors $\boldsymbol{v}$, it holds for all vectors $\boldsymbol{v}' \in \mathbb{R}^d$ by scaling. It is therefore sufficient to show that for all unit vectors $\boldsymbol{v}$,

$$(1-\epsilon) \leq \|\boldsymbol{S}\boldsymbol{U}\boldsymbol{v}\|_2^2 \leq (1+\epsilon).$$

Let $\lambda_{\min}(\boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U})$ and $\lambda_{\max}(\boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U})$ denote the minimum and maximum eigenvalues of the matrix $\boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U}$ respectively. The extreme eigenvalues are the solutions to the optimisation problems over unit vectors $\boldsymbol{v}$,

$$\lambda_{\min}(\boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U}) = \min_{\boldsymbol{v}} \|\boldsymbol{S}\boldsymbol{U}\boldsymbol{v}\|_2^2, \qquad \lambda_{\max}(\boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U}) = \max_{\boldsymbol{v}} \|\boldsymbol{S}\boldsymbol{U}\boldsymbol{v}\|_2^2.$$

The extreme eigenvalues give upper and lower bounds on the distortion in norms caused by $\boldsymbol{S}$. As we need $(1-\epsilon) \leq \|\boldsymbol{S}\boldsymbol{U}\boldsymbol{v}\|_2^2 \leq (1+\epsilon)$, the matrix $\boldsymbol{S}$ is then an ϵ-subspace embedding if and only if $|1-\lambda_{\min}(\boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U})| \leq \epsilon$ and $|1-\lambda_{\max}(\boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U})| \leq \epsilon$. The singular values of the matrix $\boldsymbol{I}_d - \boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U}$ are given by the absolute values of the eigenvalues of $\boldsymbol{I}_d - \boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U}$. The maximal singular value is then

$$\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U}) = \max(|1-\lambda_{\min}(\boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U})|, |1-\lambda_{\max}(\boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U})|) \tag{6.6}$$

It follows that $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U}) \leq \epsilon$ is a necessary and sufficient condition for $|1-\lambda_{\min}(\boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U})| \leq \epsilon$ and $|1-\lambda_{\max}(\boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U})| \leq \epsilon$ to hold jointly. As such the matrix $\boldsymbol{S}$ is an ϵ-subspace embedding if and only if $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U}) \leq \epsilon$ □

As discussed in Woodruff (2014, Chapter 2), the bounds in Table 6.1 can be obtained by giving upper bounds on the failure probability

$$1 - \Pr(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}) = \Pr(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U}) > \epsilon).$$

Let $\|\boldsymbol{A}\|_F$ denote the Frobenius norm of a $n \times d$ matrix $\boldsymbol{A}$,

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{d} A_{ij}^2}.$$

The proof for the Clarkson-Woodruff sketch uses Markov's inequality and the fact that $\sigma_{\max}^2(\boldsymbol{A}) \leq \|\boldsymbol{A}\|_F^2$ (Nelson and Nguyên, 2013; Meng, 2014). The key argument is,

$$\begin{aligned}
\Pr(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U}) > \epsilon) &= \Pr(\sigma_{\max}^2(\boldsymbol{I}_d - \boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U}) > \epsilon^2) \\
&\leq \epsilon^{-2}\mathbb{E}\left(\sigma_{\max}^2(\boldsymbol{I}_d - \boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U})\right]\right) \\
&\leq \epsilon^{-2}\mathbb{E}\left(\|\boldsymbol{I}_d - \boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U}\|_F^2\right).
\end{aligned}$$

The expected squared Frobenius norm can be upper bounded using properties of the Clarkson-Woodruff sketch and the fact that $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_d$ (Nelson and Nguyên, 2013; Meng, 2014).

The proof for the Hadamard sketch uses a matrix Chernoff bound and Boole's inequality to upper bound $\Pr(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U}) > \epsilon)$ (Tropp, 2011; Meng, 2014). The matrix Chernoff bound is used to bound the extreme eigenvalues of a random matrix. Boole's inequality upper bounds the probability of a union of events, $\Pr\left(\bigcup_{i=1}^m B_i\right) \leq \sum_{i=1}^m \Pr(B_i)$ for any countable set of events $B_1, \ldots, B_m$. The argument is more technical and not reviewed here. See Tropp (2011) for an insightful proof.

The different functional forms of the bounds in Table 6.1 are due to the different methods of proof, and the different statistical properties of the Gaussian, Hadamard and the Clarkson-Woodruff sketches. The use of chained bounds makes it difficult to keep track of the necessary constant factors to obtain a tight bound on the embedding probability. Finite sample, worst case bounds can be pessimistic in regards to the performance of the randomised algorithm (Halko et al., 2011; Mahoney and Drineas, 2016). We propose to use asymptotic random matrix theory to characterise the distribution of the critical term $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U})$. The singular values of large random matrices is a well studied topic in random matrix theory, and we can draw from existing results to analyse the behaviour of sketching algorithms (Bai et al., 2014). Given the asymptotic distribution of $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{U})$ we can give the asymptotic probability of obtaining an ϵ-subspace embedding rather than forming a lower bound. We first analyse the Gaussian sketch.

6.5.2 Gaussian sketch

From Lemma 6.2, the embedding probability is a function of the maximum singular value

$$\Pr(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}) = \Pr(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}) \leq \epsilon),$$

As noted in Meng (2014, Section 2.3), when using a Gaussian sketch it is instructive to consider directly the distribution of the random variable $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})$. Lemma 6.2 helps to show why the Gaussian projection is useful as a data oblivious sketch. Consider an arbitrary $n \times d$ data matrix $\boldsymbol{A}$. Let the singular value decomposition of $\boldsymbol{A}$ be given by $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\mathsf{T}$. The success probability of interest is can be expressed as

$$\Pr(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}) = \Pr(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}) \leq \epsilon).$$

Note that on the right hand side the sketching matrix only enters through the term $\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}$. As $\boldsymbol{S}$ is a matrix of independent Gaussians with mean zero and variance $1/k$, it is possible to show that

$$\begin{aligned}\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U} &\sim \text{Wishart}\left(k, \boldsymbol{U}^\mathsf{T}\boldsymbol{U}/k\right) \\ &= \text{Wishart}\left(k, \boldsymbol{I}_d/k\right),\end{aligned}$$

where the second line follows as $\boldsymbol{U}$ is an orthonormal matrix. The key term $\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}$ is in some sense a pivotal quantity, as its distribution is invariant to the actual values of the data matrix $\boldsymbol{A}$. When using a Gaussian sketch, the probability of obtaining an ϵ-subspace embedding has no dependence on the number of original observations n, or on the values in the data matrix $\boldsymbol{A}$. This is a useful property for a data oblivious sketch, as it is possible to develop universal performance guarantees that will hold for any possible source dataset. This invariance property is also noted in Meng (2014), although the derivation is different.

Let us define the random matrix $\boldsymbol{W} \sim \text{Wishart}(k, \boldsymbol{I}_d/k)$. The success probability of interest can then be expressed in terms of the Wishart distribution,

$$\begin{aligned}\Pr(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}) &= \Pr\left(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}) \leq \epsilon\right) \\ &= \Pr(\sigma_{\max}\left(\boldsymbol{I}_d - \boldsymbol{W}\right) \leq \epsilon). \qquad (6.7)\end{aligned}$$

Now as $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W}) = \max(|1 - \lambda_{\min}(\boldsymbol{W})|, |1 - \lambda_{\max}(\boldsymbol{W})|)$ the embedding probability can be expressed in terms of the extreme eigenvalues of the Wishart distribution. The embedding probability of interest has the representation

$$\begin{aligned}\Pr(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}) &= \Pr(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W}) \leq \epsilon) \\ &= \Pr\left(|1 - \lambda_{\min}(\boldsymbol{W})| \leq \epsilon, |1 - \lambda_{\max}(\boldsymbol{W})| \leq \epsilon\right), \qquad (6.8)\end{aligned}$$

where $\boldsymbol{W} \sim \text{Wishart}(k, \boldsymbol{I}_d/k)$. It is difficult to obtain a closed form expression for the embedding probability as it involves the joint distribution of the extreme eigenvalues. Meng forms a lower bound on this probability using concentration results on the eigenvalues of the Wishart distribution. The lower bound is reported here in Lemma 6.3.

Lemma 6.3. *(Meng, 2014, Lemma 11) Suppose we have an arbitrary $n \times d$ data matrix $\boldsymbol{A}$ where $n > d$ and $\boldsymbol{A}$ is of rank d. Suppose we take a Gaussian projection with $k \geq 6(\sqrt{d} + \sqrt{2}\log(2/\delta))^2/\epsilon^2$ then*

$$\Pr(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}) \geq 1 - \delta.$$

For a proof see (Meng, 2014, Chapter 2, Section 2.5). The bound can be inverted to give a lower bound on the success probability as a function of ϵ. Doing so gives

$$\Pr(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}) \geq 1 - \exp\left[\log(2) - \frac{(\sqrt{k\epsilon^2/6} - \sqrt{d})}{\sqrt{2}}\right]. \qquad (6.9)$$

We can investigate the tightness of the bound (6.9) through simulation. From equation (6.7), for an arbitrary $n \times d$ data matrix $\boldsymbol{A}$,

$$\Pr(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}) = \Pr(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W}) < \epsilon),$$

where $\boldsymbol{W} \sim \text{Wishart}(k, \boldsymbol{I}_d/k)$. To estimate the embedding probability for the Gaussian sketch we can simulate $\boldsymbol{W} \sim \text{Wishart}(k, \boldsymbol{I}_d/k)$ and look at the empirical distribution of the random variable $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W})$. We first generated $B = 10000$ random Wishart matrices $\boldsymbol{W}^{[1]}, \ldots, \boldsymbol{W}^{[B]}$. For each simulated matrix $\boldsymbol{W}^{[b]}$ we computed the distortion factor $\epsilon^{[b]}$:

$$\epsilon^{[b]} = \sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W}^{[b]}), \tag{6.10}$$

for $b = 1, \ldots, B$. The simulated distortion factors $\epsilon^{[1]}, \ldots, \epsilon^{[1B]}$ can be used to give a Monte Carlo estimate of the embedding probability:

$$\begin{aligned}\widehat{\Pr}(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}) &= \widehat{\Pr}(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W}) \leq \epsilon) \\ &= \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}(\epsilon^{[b]} \leq \epsilon).\end{aligned}$$

We used the `ARPACK` library Lehoucq et al. (1998) to compute the maximum singular value of the high-dimensional matrices. The estimated embedding probabilities are displayed in Figure 6.1 for different dimensions d. The sketch size was kept at $k = 20 \times d$. The red line shows the empirical probability of obtaining an ϵ-subspace embedding. The dot-dash line shows the lower bound of Meng (2014), given in equation (6.9). The solid vertical line gives an asymptotic limiting value that will be discussed later in section 6.7.1.

Comparing the empirical embedding probabilities to the lower bound, we see that the lower bound is quite loose. The lower bound is zero in each plot at points where the empirical cdf is close to one. To obtain a tighter prediction on the embedding probability we consider the asymptotic distribution of the eigenvalues of a random Wishart matrix. This is a well studied area of random matrix theory (Edelman, 1988). Asymptotic analysis is adopted in order to obtain point estimate of the success probability rather than a worst case bound. We develop two asymptotic expressions for the embedding probability. We introduce the random matrix asymptotics by first considering the pointwise limit of the extreme eigenvalues. The pointwise asymptotic result suggests a sharp phase change in the success probability of the algorithm. We then develop a more accurate approximation for the embedding probability by making a connection to the Tracy-Widom distribution.

6.6 Asymptotics

Our asymptotic arguments in section 6.7 concern the convergence of probability measures. Billingsley (1999) is an authoritative reference on the topic. We now recap some useful foundational theory, as is presented in Van Der Vaart (1998, Chapter 2). Let $(\boldsymbol{Z}_n)_{n\in\mathbb{N}}$ denote a sequence of real v-dimensional random vectors with cumulative distribution functions $(F_n)_{n\in\mathbb{N}}$. Let Z denote a real v-dimensional random vector with cumulative distribution function F. The sequence of random vectors $(\boldsymbol{Z}_n)$ converges in distribution to $\boldsymbol{Z}$ if

$$\lim_{n\to\infty} F_n(\boldsymbol{z}) = F(\boldsymbol{z}),$$

at every point $\boldsymbol{z} \in \mathbb{R}^v$ where the limit distribution F is continuous. Convergence in distribution is denoted as $\boldsymbol{Z}_n \xrightarrow{d} \boldsymbol{Z}$. Convergence in distribution is often referred to as weak convergence. Let $\|.\|_2$ denote the Euclidean norm. A sequence of random vectors $(\boldsymbol{Z}_n)_{n\in\mathbb{N}}$, is said to converge in probability to a random vector $\boldsymbol{Z}$ if for all $\delta > 0$:

$$\lim_{n\to\infty} \Pr(\|\boldsymbol{Z}_n - \boldsymbol{Z}\|_2 > \delta) = 0.$$

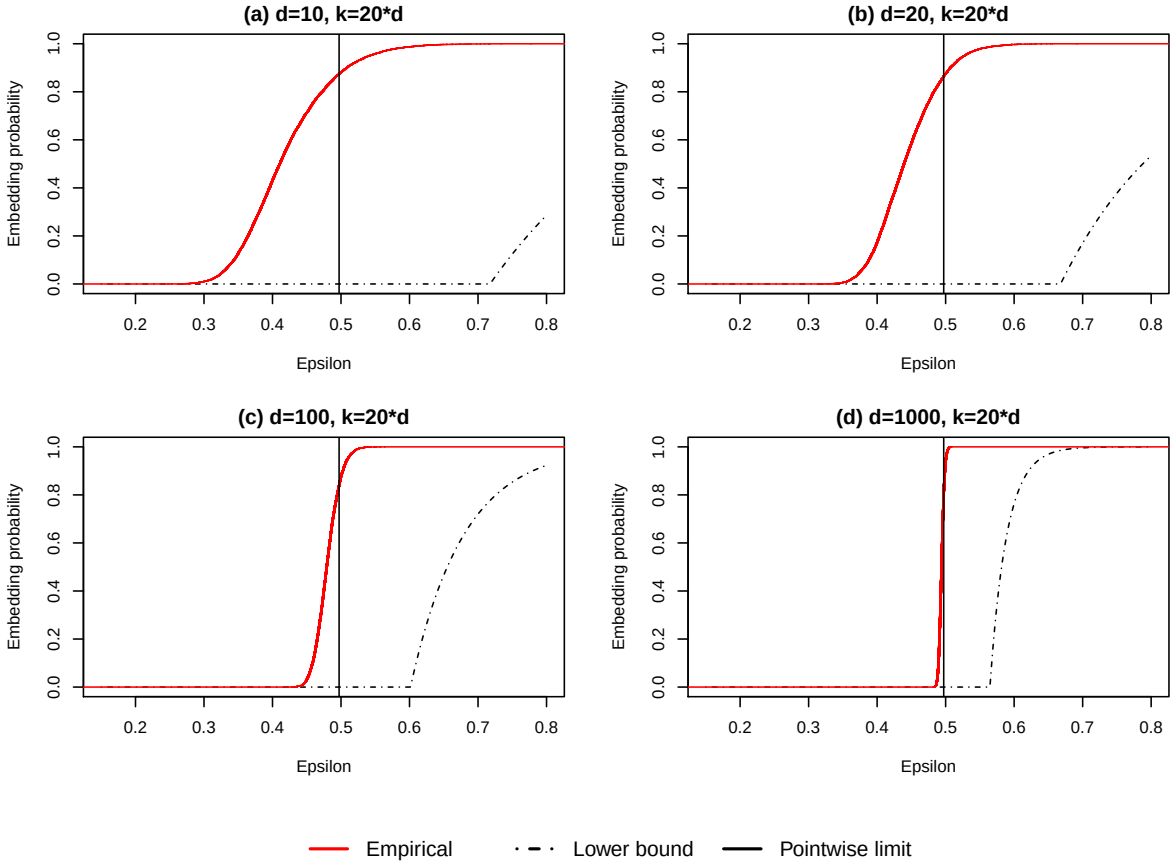


Figure 6.1: Comparison of simulated embedding probabilities against theoretical results at different k and d. The sketch size to variables ratio is kept constant at twenty. We want the sketching matrix $\boldsymbol{S}$ to be an ϵ-subspace embedding for the source dataset with small ϵ with high probability. The y-axis gives the proportion of times we attain an ϵ-subspace embedding. The x-axis gives the distortion factor ϵ (recall Definition 6.1). The dot-dash line gives a finite sample lower bound. The vertical line gives the asymptotic limiting distribution as d increases. As d increases the empirical cdf concentrates around the step-function given by pointwise asymptotic theory (Theorem 6.5).

Convergence in probability is denoted $\boldsymbol{Z}_n \xrightarrow{p} \boldsymbol{Z}$. Convergence in probability requires that $(\boldsymbol{Z}_n)$ and $\boldsymbol{Z}$ be defined on the same probability space. Convergence in distribution has no such requirement.

The Portmanteau lemma gives a number of useful equivalent definitions of convergence in distribution (weak convergence).

Lemma 6.4 (Portmanteau)**.** *Let $(Z_n)_{n\in\mathbb{N}}$ denote a sequence of random vectors of fixed dimension, and $\boldsymbol{Z}$ denote another random vector of the same dimension. The following statements are equivalent, where limits are being taken in n:*

(a) $\Pr(\boldsymbol{Z}_n \leq \boldsymbol{z}) \to \Pr(\boldsymbol{Z} \leq \boldsymbol{x})$ *at all continuity points $\boldsymbol{z}$ of the cumulative distribution function* $\Pr(\boldsymbol{Z} \leq \boldsymbol{z})$.

(b) $\mathbb{E}f(\boldsymbol{Z}_n) \to \mathbb{E}f(\boldsymbol{Z})$ *for all bounded, continuous function f.*

(c) $\mathbb{E}f(\boldsymbol{Z}_n) \to \mathbb{E}f(\boldsymbol{Z})$ *for all bounded Lipschitz functions f*

(d) $\liminf \mathbb{E}f(\boldsymbol{Z}_n) \geq \mathbb{E}f(\boldsymbol{Z})$ *for all nonnegative, continuous functions f.*

(e) $\liminf \Pr(\boldsymbol{Z}_n \in G) \geq \Pr(\boldsymbol{Z} \in G)$ *for every open set G.*

(f) $\limsup \Pr(\boldsymbol{Z}_n \in F) \leq \Pr(\boldsymbol{Z} \in F)$ *for every closed set F.*

(g) $\Pr(\boldsymbol{Z}_n \in B) \to \Pr(\boldsymbol{Z} \in B)$ *for all Borel sets* B *with* $\Pr(\boldsymbol{X} \in \partial B) = 0$*, where* ∂B *denotes the boundary of the set* B*. The boundary is defined as the closure of the set* B *minus the interior of* B*, so* $\partial B = \overline{B} \setminus B^o$*.*

The continuous mapping theorem is a useful result for studying the weak convergence of functions of random variables. The continuous mapping theorem is quite powerful as we only need to show that the limiting random variable $\boldsymbol{Z}$ satisfies $P(\boldsymbol{Z} \in C) = 1$. Possible discontinuities of $g(\boldsymbol{Z}_n)$ do not cause any difficulties in establishing a limit theorem.

Theorem 6.1 (Continuous Mapping Theorem)**.** *Let* $(\boldsymbol{Z}_n)_{n\in\mathbb{N}}$ *indicate a sequence of* v*-dimensional random vectors. Let* $\boldsymbol{Z}$ *denote another random vector of dimension* v*. Suppose the function* $g : \mathbb{R}^v \to \mathbb{R}^m$ *is continuous at every point of a set* C *such that* $P(\boldsymbol{Z} \in C) = 1$*. Then the following results hold*

(a) If $\boldsymbol{Z}_n \xrightarrow{d} \boldsymbol{Z}$ *then* $g(\boldsymbol{Z}_n) \xrightarrow{d} g(\boldsymbol{Z})$*.*

(b) If $\boldsymbol{Z}_n \xrightarrow{p} \boldsymbol{Z}$ *then* $g(\boldsymbol{Z}_n) \xrightarrow{p} g(\boldsymbol{Z})$*.*

There are some useful relationships between convergence in probability and convergence in distribution. Additionally, there are some useful relationships between joint and marginal convergence when one or more sequences in question converges weakly to a constant. The following theorem summarises some useful results.

Theorem 6.2 (Relationship between modes of convergence)**.** *Let* $(\boldsymbol{Z}_n)_{n\in\mathbb{N}}$ *and* $(\boldsymbol{Y}_n)_{n\in\mathbb{N}}$ *be sequences of random vectors. Let* $\boldsymbol{Z}$ *and* $\boldsymbol{Y}$ *denote random vectors. Then the following relationships hold:*

(a) $\boldsymbol{X}_n \xrightarrow{p} \boldsymbol{X}$ *implies* $\boldsymbol{X}_n \xrightarrow{d} \boldsymbol{X}$*.*

(a) $\boldsymbol{X}_n \xrightarrow{p} \boldsymbol{c}$ *for a constant* $\boldsymbol{c}$ *if and only if* $\boldsymbol{X}_n \xrightarrow{d} \boldsymbol{c}$*.*

(a) If $\boldsymbol{X}_n \xrightarrow{p} \boldsymbol{c}_1$ *and* $\boldsymbol{Y}_n \xrightarrow{p} \boldsymbol{c}_2$ *for constants* $\boldsymbol{c}_1$ *and* $\boldsymbol{c}_2$*, then* $(\boldsymbol{X}_n, \boldsymbol{Y}_n) \xrightarrow{p} (\boldsymbol{c}_1, \boldsymbol{c}_2)$

Slutsky's theorem concerns the joint convergence of functions of random variables when there is some convergence to a constant.

Theorem 6.3 (Slutsky)**.** *Let* $(\boldsymbol{Z}_n)_{n\in\mathbb{N}}$ *and* $(\boldsymbol{Y}_n)_{n\in\mathbb{N}}$ *be sequences of random vectors or random variables. Let* $\boldsymbol{X}$ *denote random vector or random variable. If* $\boldsymbol{X}_n \xrightarrow{d} \boldsymbol{X}$ *and* $\boldsymbol{Y}_n \xrightarrow{d} \boldsymbol{c}$ *for a constant* $\boldsymbol{c}$*, then*

(a) $\boldsymbol{X}_n + \boldsymbol{Y}_n \xrightarrow{d} \boldsymbol{X} + c$

(a) $\boldsymbol{Y}_n \boldsymbol{X}_n \xrightarrow{d} \boldsymbol{c}\boldsymbol{X}$

(a) $\boldsymbol{Y}_n^{-1} \boldsymbol{X}_n \xrightarrow{d} \boldsymbol{c}^{-1}\boldsymbol{X}$ *provided that* $\boldsymbol{c} \neq 0$*.*

Lemma 6.5 (Uniform convergence)**.** *Suppose that* $(\boldsymbol{Z}_n)$*converges in distribution to a random vector* $\boldsymbol{Z}$ *with a continuous distribution function. Then*

$$\lim_{n\to\infty} \sup_{\boldsymbol{z}} |\Pr(\boldsymbol{Z}_n \leq \boldsymbol{z}) - \Pr(\boldsymbol{Z} \leq \boldsymbol{z})| = 0.$$

Proofs of all the results in this subsection are given in Chapter 2 of Van Der Vaart (1998).

6.7 Random matrix theory

6.7.1 Pointwise limit

The extreme eigenvalues of a Wishart random matrix converge in probability to fixed values as both the dimension and degrees of freedom expand. The result for the largest eigenvalue is due to Geman (1980) and the result for the smallest eigenvalue is due to Silverstein (1985).

Theorem 6.4. *(Geman, 1980; Silverstein, 1985)*
Consider a sequence of $\text{Wishart}(k, \boldsymbol{I}_d/k)$ *random matrices where the degrees of freedom* k *and dimension* d *are both taken to infinity. Suppose that the variables to samples ratio* d/k *converges to a constant* $(d/k) \to \alpha$*, where* $\alpha \in (0, 1]$*. Then the extreme eigenvalues of the random matrix,* $\lambda_{\min}$ *and* $\lambda_{\max}$ *converge in probability to the limits*

$$(i)\ \lambda_{\min} \xrightarrow{p} (1 - \sqrt{\alpha})^2, \tag{6.11}$$

$$(ii)\ \lambda_{\max} \xrightarrow{p} (1 + \sqrt{\alpha})^2. \tag{6.12}$$

Theorem 6.4 and the continuous mapping theorem (Theorem 6.1) can be used to determine the asymptotic embedding probability for the Gaussian sketch. Theorem 6.5 gives the limiting probability of obtaining an ϵ-subspace embedding for a Gaussian sketch. The limit is asymptotic in n, d and k. This can be interpreted as a type of Big Data asymptotic, where we consider tall and wide datasets through the limit in n and d, and increasing sketch sizes k to cope with the with the expanding number of variables d. In Theorem 6.5 the limiting variables to sketch ratio d/k is taken to a constant that is less than one. This is reasonable as the sketched dataset will only have a full rank covariance matrix if $k > d$.

Theorem 6.5. *Suppose we have an arbitrary* $n \times d$ *data matrix* $\boldsymbol{A}$ *where* $n > d$ *and* $\boldsymbol{A}$ *is of rank* d*. Assume we take a Gaussian sketch of size* k*. Then asymptotically in* n, k *and* d*, with* $d/k \to \alpha$ *where* $\alpha \in (0, 1]$*,*

$$\lim_{n,d,k\to\infty} \Pr(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}) = \begin{cases} 0 & \text{if } \epsilon < (1 + \sqrt{\alpha})^2 - 1 \\ 1 & \text{if } \epsilon > (1 + \sqrt{\alpha})^2 - 1 \end{cases}$$

Before giving a proof we first make a comment on the practical interpretation of Theorem 6.5. For large d and k, we expect the embedding probability to be a step function at $(1 + \sqrt{d/k})^2 - 1$. In the simulations depicted in Figure 6.1, the sketch size to variable ratio was held constant at twenty. As k and d grow larger, the embedding probability curve should concentrate at $(1 + \sqrt{1/20})^2 - 1 \approx 0.5$. This value is plotted as a dashed line in each panel. Moving through the panels in the order (a) through (d) we can see visible convergence to the pointwise limit around 0.5. As d and k increase, the empirical cdf of the simulated values $\epsilon^{[1]}, \ldots, \epsilon^{[B]}$, (see equation (6.10)) approaches a step function at the pointwise limit.

Proof: Let $\boldsymbol{W} \sim \text{Wishart}(k, \boldsymbol{I}_d/k)$, and let $\lambda_{\min}$ and $\lambda_{\max}$ denote the minimum and maximum eigenvalues of $\boldsymbol{W}$ respectively. Using Slutsky's theorem and the continuous mapping theorem we have the joint convergence result

$$\begin{bmatrix} |1 - \lambda_{\min}| \\ |1 - \lambda_{\max}| \end{bmatrix} \xrightarrow{p} \begin{bmatrix} |1 - (1 - \sqrt{\alpha})^2| \\ |1 - (1 + \sqrt{\alpha})^2| \end{bmatrix}. \tag{6.13}$$

For large k and d, the maximum eigenvalue $\lambda_{\max}$ is expected to show greater deviation from one than the minimum eigenvalue $\lambda_{\min}$. Over the interval $\alpha \in (0, 1]$ it holds that

$$|1 - (1 + \sqrt{\alpha})^2| > |1 - (1 - \sqrt{\alpha})^2|.$$

Applying the continuous mapping theorem to the random vector in (6.13),

$$\max \begin{bmatrix} |1 - \lambda_{\min}| \\ |1 - \lambda_{\max}| \end{bmatrix} \xrightarrow{p} \max \begin{bmatrix} |1 - (1 - \sqrt{\alpha})^2| \\ |1 - (1 + \sqrt{\alpha})^2| \end{bmatrix},$$

yielding $\max(|1-\lambda_{\min}|, |1-\lambda_{\max}|) \xrightarrow{p} |1-(1+\sqrt{\alpha})^2|$. Now as $(1+\sqrt{\alpha})^2$ is greater than one for all $\alpha > 0$, the absolute value sign can be removed in the limit giving the equivalent statement $\max(|1-\lambda_{\min}|, |1-\lambda_{\max}|) \xrightarrow{p} (1+\sqrt{\alpha})^2 - 1$. Recalling that $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W}) = \max(|1-\lambda_{\min}|, |1-\lambda_{\max}|)$, we establish convergence of the limiting singular value

$$\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W}) \xrightarrow{p} (1+\sqrt{\alpha})^2 - 1. \tag{6.14}$$

Property Portmanteau lemma then gives the probabilistic statement

$$\lim_{n,d,k\to\infty} \Pr(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W}) \leq \epsilon) = \begin{cases} 0 & \text{if } \epsilon < (1+\sqrt{\alpha})^2 - 1, \\ 1 & \text{if } \epsilon > (1+\sqrt{\alpha})^2 - 1. \end{cases}$$

As $\epsilon = (1+\sqrt{\alpha})^2 - 1$ is a discontinuity point of the limiting distribution function we do not make a statement about the case $\epsilon = (1+\sqrt{\alpha})^2 - 1$. From (6.7) we have the equality in limits

$$\lim_{n,d,k\to\infty} \Pr(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}) = \lim_{n,d,k\to\infty} \Pr(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W}) \leq \epsilon),$$

giving the final result. □

In Figure 6.1 the Monte Carlo estimate of the embedding probability curve shows deviation from the sharp step function predicted by Theorem 6.5. We can obtain a more accurate approximation of the embedding probability using more sophisticated asymptotics.

6.7.2 Tracy-Widom limit

We will show that the embedding probability (6.1) under the Gaussian sketch is closely related to the Tracy-Widom law. The Tracy-Widom distribution is the limiting distribution of the maximum eigenvalue of a random Wishart$(m, \boldsymbol{I}_m)$ matrix. Note that in the previous expression the degrees of freedom m matches the dimension of the $m \times m$ random matrix. The cumulative distribution function of the Tracy-Widom distribution $F_1(x)$, is defined as

$$F_1(x) = \exp\left(-\frac{1}{2}\int_x^\infty q(t) + (t-x)q^2(t)\ dt\right).$$

Where $q(x)$ satisfies the nonlinear differential equation

$$q''(x) = xq(x) + 2q^3(x),$$

subject to the asymptotic boundary condition, $q(x) \sim \text{Ai}(x) \quad \text{as } x \to \infty$. The function $\text{Ai}(x)$ denotes the Airy function, defined as

$$\text{Ai}(x) = \frac{1}{\pi}\int_0^\infty \cos\left(\frac{t^3}{3} + xt\right)\ dt.$$

The Airy equation $y'' - xy = 0$ has solution $y = \text{Ai}(x)$ under the boundary condition $y \to 0$ as $x \to \infty$. Figure 6.2 plots the density function of the Tracy-Widom distribution and the Airy function. The R package `RMTstat` has a suite of functions for working with the Tracy-Widom distribution (Johnstone et al., 2014). The Tracy-Widom law describes Wishart$(m, \boldsymbol{I}_m)$ random matrices, however we are interested in situations where the degrees of freedom may not match the dimension of the matrix. For sketching algorithms the random matrix of interest $\boldsymbol{W} \sim \text{Wishart}(k, \boldsymbol{I}_d/k)$ has the degrees of freedom controlled by the sketch size k, and the dimension of the matrix is given by the number of variables in the source dataset d. Johnstone (2001) showed that Tracy-Widom law also gives the asymptotic distribution of the maximum eigenvalue of a Wishart$(k, \boldsymbol{I}_d/k)$ matrix after appropriate centring and scaling. In subsequent work Ma (2012) showed that the rate of convergence could be improved by using different centering and scaling constants than in Johnstone (2001). We present the convergence result given by Ma.

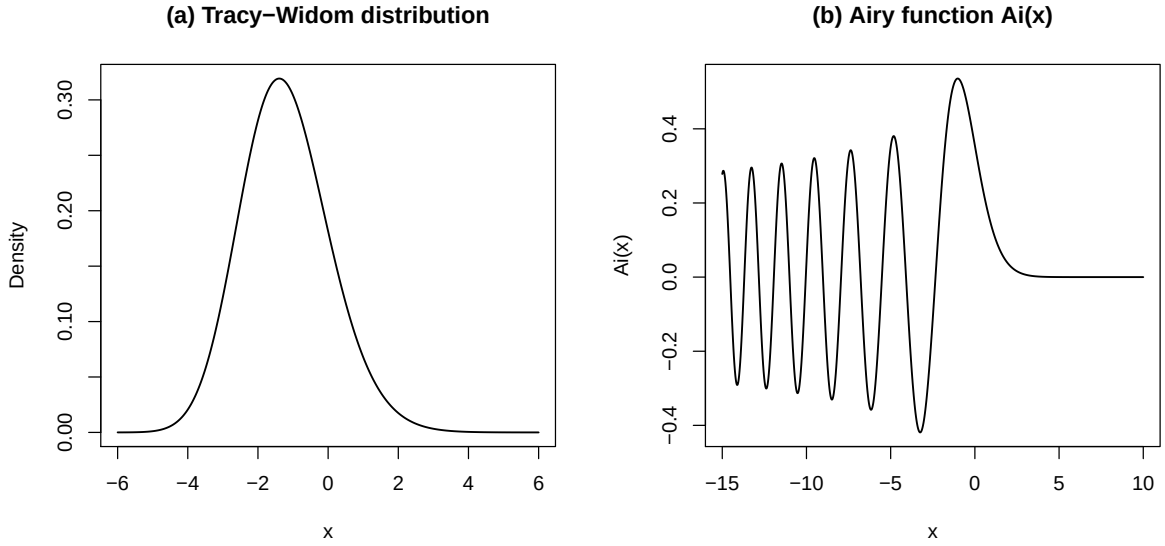


Figure 6.2: (a) Tracy-Widom distribution. (b) Airy function. The Tracy-Widom distribution has a significant role in describing the asymptotic distributions of the eigenvalues of large random matrices.

Theorem 6.6. *(Ma, 2012)*
Consider a sequence of $\text{Wishart}(k, \boldsymbol{I}_d/k)$ *random matrices where the degrees of freedom* k *and dimension* d *are both taken to infinity. Let* $\lambda_{\max}$ *denote the maximum eigenvalue of the random matrix. Suppose that the variables to samples ratio* d/k *converges such that* $d/k \to \alpha$ *where* $\alpha \in (0, 1]$*. Define the centring and scaling constants as*

$$\mu_{k,d} = k^{-1}(\sqrt{k - 1/2} + \sqrt{d - 1/2})^2,$$
$$\sigma_{k,d} = k^{-1}(\sqrt{k - 1/2} + \sqrt{d - 1/2}) \left(\frac{1}{\sqrt{k - 1/2}} + \frac{1}{\sqrt{d - 1/2}} \right)^{1/3}.$$

Then

$$\frac{(\lambda_{\max} - \mu_{k,d})}{\sigma_{k,d}} \xrightarrow{d} Z,$$

where $Z \sim F_1$ *and* F_1 *is the Tracy-Widom distribution.*

The limiting values of $\mu_{k,d}$ and $\sigma_{k,d}$ are $(1 + \sqrt{\alpha})^2$ and 0 as we take k, d to infinity with $d/k \to \alpha$. This shows a correspondence with the pointwise limit of maximum eigenvalue given in Theorem 6.4. Theorem 6.7 uses the Tracy-Widom law to describe the asymptotic embedding probability when using a Gaussian sketch.

Theorem 6.7. *Suppose we have an arbitrary* $n \times d$ *data matrix* $\boldsymbol{A}$ *where* $n > d$ *and* $\boldsymbol{A}$ *is of rank* d*. Let the singular value decomposition of* $\boldsymbol{A}$ *be given by* $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\mathsf{T}$*. Furthermore assume we take a Gaussian sketch of size* k*. Consider the limit in* n, k *and* d*, such that* $d/k \to \alpha$ *with* $\alpha \in (0, 1]$*. Let* $\mu_{k,d}$ *and* $\sigma_{k,d}$ *be the centering and scaling constants given in Theorem 6.6. The asymptotically in* n, d *and* k*,*

$$(i) \lim_{n,d,k \to \infty} \left| \Pr\left(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}\right) - \Pr\left(Z \leq \frac{\epsilon + 1 - \mu_{k,d}}{\sigma_{k,d}} \right) \right| = 0.$$

Where $Z \sim F_1$ *and* F_1 *is the Tracy-Widom distribution. Furthermore we have convergence in distribution*

$$(ii) \quad \frac{\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}) - \mu_{k,d} + 1}{\sigma_{k,d}} \xrightarrow{d} Z.$$

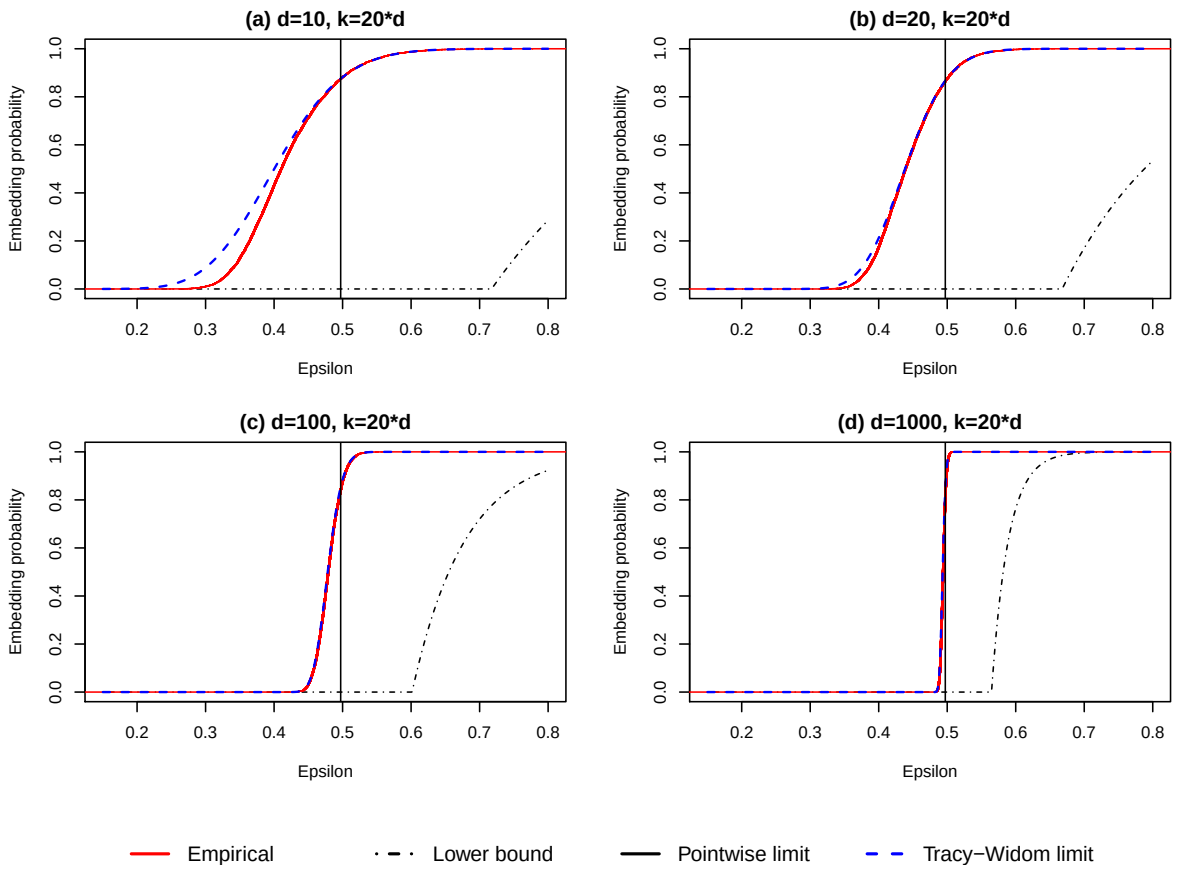


Figure 6.3: Empirical probability of obtaining an ϵ-subspace embedding at different k and d. The sketch to variables ratio is kept constant at twenty. We want the sketching matrix $\boldsymbol{S}$ to be an ϵ-subspace embedding for the source dataset with small ϵ with high probability. The y-axis gives the proportion of times we attain an ϵ-subspace embedding. The x-axis gives the distortion factor ϵ (recall Definition 6.1). The empirical cdf is obtained by simulating $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W})$ where $\boldsymbol{W} \sim \text{Wishart}(k, \boldsymbol{I}_d/k)$. The Tracy-Widom law gives the most accurate predictions of the embedding probability in the simulation. The lower bound is conservative at each d. The Tracy-Widom based prediction (Theorem 6.7) describes the fluctuations around the sharp step-function given by the pointwise asymptotic theory (Theorem 6.5). The accuracy of the Tracy-Widom approximation improves as d increases.

Before presenting the proof we show an application of Theorem 6.7. Result (i) gives an asymptotic approximation for the embedding probability given that we take a sketch of size k of a source dataset with d variables. We compare the Tracy-Widom approximation in Theorem 6.7 to the simulation results reported in section 6.5.2. Figure 6.3 displays the results. The asymptotic expression for the embedding probability is superimposed as a blue dashed line over the empirical estimate (red line). The pointwise limit discussed in section 6.7.1 is again plotted as a solid vertical line. The Tracy-Widom limit is clearly more accurate than the pointwise approximation, modelling the fluctuation around the limiting value $(1 - \sqrt{(1/20)})^2 - 1 \approx 0.5$. The Tracy-Widom approximation improves as k and d grow larger. In (a) we can see that the left tail of the Tracy-Widom approximation is not accurate for $d = 10$. In (c) we can see that the approximation is very good at $d = 100$. The lower bound (6.9) is again plotted as the dot-dash line. Although an asymptotic result, the Tracy-Widom limit gives a more accurate prediction of the embedding probability than the finite sample lower bound in this simulation.

Theorem 6.7 (ii) gives the asymptotic distribution of $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})$ for an arbitrary data matrix $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\mathsf{T}$. The simulated values $\epsilon^{[1]}, \ldots, \epsilon^{[B]}$ (recall equation (6.10)) can be used to estimate the density of $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})$. Figure 6.4 compares the asymptotic theoretical density of $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})$ to a kernel density estimate from the simulated values. The agreement improves as d grows, with the asymptotic approximation being extremely good for $d = 100$ and $d = 1000$. Theorem 6.7 is not an immediate extension of Theorem 6.6 as Theorem 6.6 does not pertain describe the joint distribution

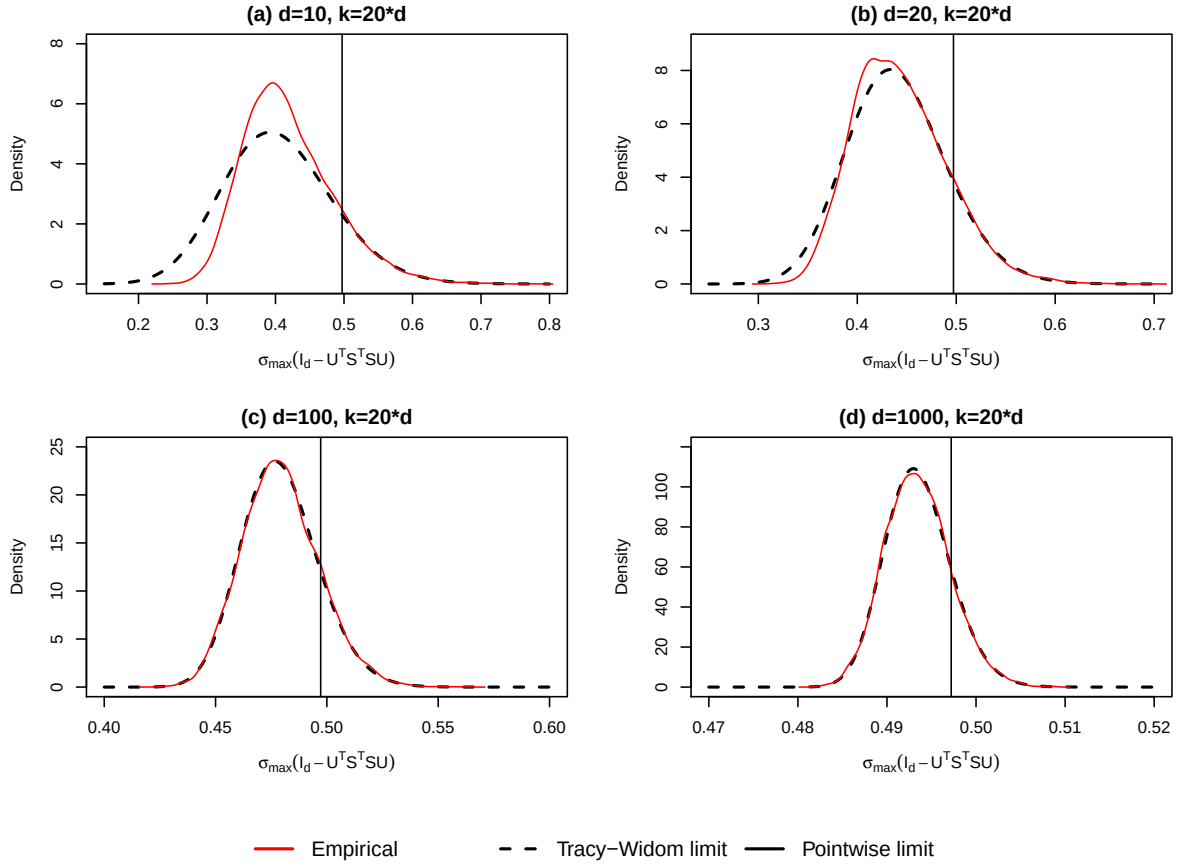


Figure 6.4: Observed and theoretical density of $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})$ at different k and d. The sketch to variables ratio is kept constant at twenty. We want the sketching matrix $\boldsymbol{S}$ to be an ϵ-subspace embedding for the source dataset with small ϵ with high probability. The sketching matrix $\boldsymbol{S}$ is an ϵ-subspace embedding for the source dataset if and only if $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}) \leq \epsilon$. The x-axis limits are different in each plot. The accuracy of the Tracy-Widom approximation improves as d increases as is expected from the asymptotic theory (Theorem 6.7).

of $\lambda_{\min}$ and $\lambda_{\max}$. The embedding probability involves the joint distribution of $\lambda_{\min}$ and $\lambda_{\max}$. We use Theorem 6.6 in conjunction with the Portmanteau lemma, Theorem 6.5 and the continuous mapping theorem in order to establish Theorem 6.7.

We now present the proof of Theorem 6.7.

Proof: Let $\boldsymbol{W} \sim \text{Wishart}(k, \boldsymbol{I}_d/k)$, and let $\lambda_{\min}$ and $\lambda_{\max}$ denote the minimum and maximum eigenvalues of $\boldsymbol{W}$ respectively. The majority of the proof comes down to showing that $\lambda_{\max}$ controls the embedding probability. Using the Portmanteau lemma (Lemma 6.4) we will show that

$$\lim_{d,k\to\infty} \Pr(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W}) \leq \epsilon) = \lim_{d,k\to\infty} \Pr\left(|1 - \lambda_{\max}| \leq \epsilon\right).$$

Recall the key expression given in (6.8),

$$\begin{aligned}\Pr(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}) &= \Pr(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W}) \leq \epsilon) \\ &= \Pr\left(|1 - \lambda_{\min}| \leq \epsilon, |1 - \lambda_{\max}| \leq \epsilon\right).\end{aligned}$$

The Tracy-Widom law describes the marginal distributions of $\lambda_{\min}$ and $\lambda_{\max}$. We would like to avoid working with the joint distribution of the extreme eigenvalues, and instead restrict attention to the distribution of the maximum. Let denote the random vector $\boldsymbol{X} = (|1 - \lambda_{\min}|, |1 - \lambda_{\max}|)^\mathsf{T}$. Figure 6.5 presents some diagrams that will be useful. We wish to know the probability that $\boldsymbol{X}$ lies in the region shaded region C in panel (a). For every $\epsilon > 0$ we have that $\Pr\left(|1 - \lambda_{\min}| \leq \epsilon, |1 - \lambda_{\max}| \leq \epsilon\right) = \Pr(\boldsymbol{X} \in C)$. The region C can be expressed as $C = M - R$ where M and R are the shaded regions in panels (b) and

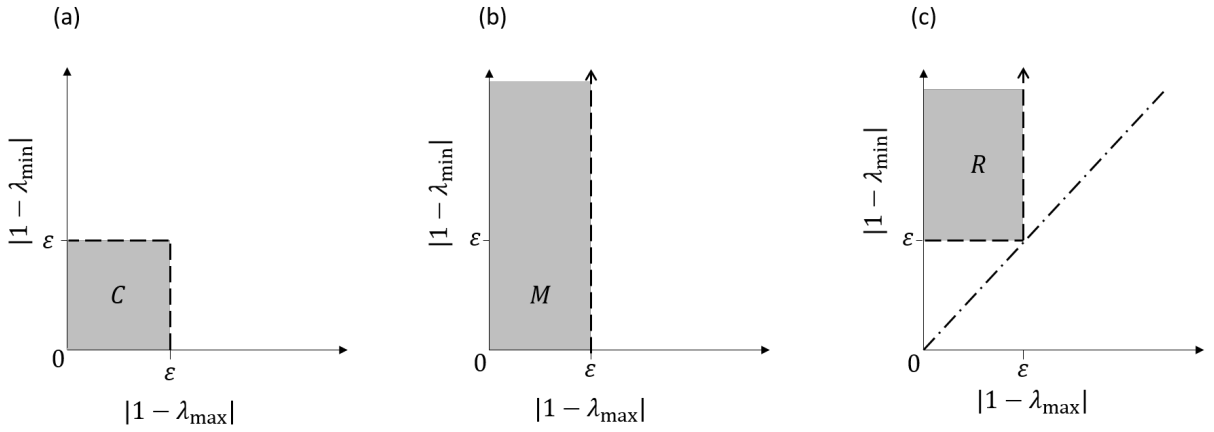


Figure 6.5: Regions of interest in determining the embedding probability. To obtain an ϵ-subspace embedding we require that $|1-\lambda_{\min}| \leq \epsilon$ and $|1-\lambda_{\max}| \leq \epsilon$. If we define $\boldsymbol{X} = (|1-\lambda_{\min}|, |1-\lambda_{\max}|)^{\mathsf{T}}$, we have that $\Pr(\boldsymbol{X} \in C) = \Pr(\boldsymbol{X} \in M) - \Pr(\boldsymbol{X} \in R)$. In panel (c) the dot-dash line gives the identity line where $|1-\lambda_{\max}| = |1-\lambda_{\min}|$.

(c) respectively. The probability $\Pr(\boldsymbol{X} \in M)$ represents the marginal probability that $|1-\lambda_{\max}| \leq \epsilon$. The probability $\Pr(\boldsymbol{X} \in R)$ represents the probability of the joint event that $(|1-\lambda_{\max}| \leq \epsilon, |1-\lambda_{\min}| > \epsilon)$. We have that

$$\Pr(\boldsymbol{X} \in C) = \Pr(\boldsymbol{X} \in M) - \Pr(\boldsymbol{X} \in R).$$

In panel (c) the dot-dash line gives the identity line where $|1-\lambda_{\max}| = |1-\lambda_{\min}|$. From Theorem 6.5 we know that as d, k tends to infinity $\boldsymbol{X}$ converges in distribution to the constant vector $\boldsymbol{X}_L = (|1-(1-\sqrt{\alpha})^2|, |1-(1+\sqrt{\alpha})^2|)^{\mathsf{T}}$. As such, asymptotically $|1-\lambda_{\max}| > |1-\lambda_{\min}|$ with probability one. Referring to panel (c), the random vector $\boldsymbol{X}_L$ takes values in the region below the dot-dash line with probability one. The limiting random vector $\boldsymbol{X}_L$ thus satisfies $\Pr(\boldsymbol{X}_L \in R) = 0$ and $\Pr(\boldsymbol{X}_L \in \partial R) = 0$. As $\boldsymbol{X} \xrightarrow{d} \boldsymbol{X}_L$, property (g) of the Portmanteau lemma (Lemma 6.4) gives that $\Pr(\boldsymbol{X} \in R) \to \Pr(\boldsymbol{X}_L \in R) = 0$. The limiting probability is then

$$\begin{aligned}
\lim_{d,k\to\infty} \Pr\left(|1-\lambda_{\min}| \leq \epsilon, |1-\lambda_{\max}| \leq \epsilon\right) &= \lim_{d,k\to\infty} \Pr(\boldsymbol{X} \in C) && (6.15)\\
&= \lim_{d,k\to\infty} \Pr(\boldsymbol{X} \in M) - \lim_{d,k\to\infty} \Pr(\boldsymbol{X} \in R) && (6.16)\\
&= \lim_{d,k\to\infty} \Pr(\boldsymbol{X} \in M) - 0 && (6.17)\\
&= \lim_{d,k\to\infty} \Pr\left(|1-\lambda_{\max}| \leq \epsilon\right).. && (6.18)
\end{aligned}$$

We have now isolated the maximum eigenvalue $\lambda_{\max}$ as the determining factor in obtaining an ϵ-subspace embedding. We make another application of the Portmanteau lemma to arrive at the final result. From here we can write

$$\Pr\left(|1-\lambda_{\max}| \leq \epsilon\right) = \Pr(\lambda_{\max} \leq \epsilon + 1) - \Pr(\lambda_{\max} \leq 1 - \epsilon). \tag{6.19}$$

From Theorem 6.5 we know that $\lambda_{\max}$ converges in distribution to the constant random variable $Z_L = (1+\sqrt{\alpha})^2$, where we have assumed $\alpha \in (0, 1]$. Let B denote the interval $(-\infty, 1]$. The limiting random variable Z_L satisfies $\Pr(Z_L \in B) = 0$ and $\Pr(Z_L \in \partial B) = 0$. As such using property g of the Portmanteau lemma, $\lim_{d,k\to\infty} \Pr(\lambda_{\max} \in B) = 0$. Now $\Pr(\lambda_{\max} \leq 1-\epsilon) \leq \Pr(\lambda_{\max} \in B)$ for any $\epsilon > 0$. We can then conclude that $\lim_{d,k\to\infty} \Pr(\lambda_{\max} \leq 1-\epsilon) = 0$ for any $\epsilon > 0$. Asymptotically, the term $\Pr(\lambda_{\max} \leq 1-\epsilon)$ drops out of the expression for the embedding probability. Taking limits over (6.19)

$$\begin{aligned}
\lim_{d,k\to\infty} \Pr\left(|1-\lambda_{\max}| \leq \epsilon\right) &= \lim_{d,k\to\infty} \Pr(\lambda_{\max} \leq \epsilon + 1) - \lim_{d,k\to\infty} \Pr(\lambda_{\max} \leq 1 - \epsilon)\\
&= \lim_{d,k\to\infty} \Pr(\lambda_{\max} \leq \epsilon + 1) - 0.
\end{aligned}$$

The asymptotic embedding probability is then related to the asymptotic cdf of $\lambda_{\max}$. Let Z be a random variable with Tracy-Widom distribution F_1. Now for any fixed d and k, it must hold that for any fixed $\epsilon > 0$,

$$\left|\Pr\left(\frac{\lambda_{\max} - \mu_{k,d}}{\sigma_{k,d}} \leq \frac{\epsilon + 1 - \mu_{k,d}}{\sigma_{k,d}}\right) - \Pr\left(Z \leq \frac{\epsilon + 1 - \mu_{k,d}}{\sigma_{k,d}}\right)\right| \leq \sup_{z \in \mathbb{R}} \left|\Pr\left(\frac{\lambda_{\max} - \mu_{k,d}}{\sigma_{k,d}} \leq z\right) - \Pr(Z \leq z)\right|.$$

Using Theorem 6.6, the random variable $(\lambda_{\max} - \mu_{k,d})/\sigma_{k,d}$ converges in distribution to the continuous random variable $Z \sim F_1$. It follows from Lemma 6.5 that

$$\lim_{d,k \to \infty} \sup_{z \in \mathbb{R}} \left|\Pr\left(\frac{\lambda_{\max} - \mu_{k,d}}{\sigma_{k,d}} \leq z\right) - \Pr(Z \leq z)\right| = 0.$$

Now by the squeeze theorem, it holds that for all $\epsilon > 0$,

$$\lim_{d,k \to \infty} \left|\Pr\left(\frac{\lambda_{\max} - \mu_{k,d}}{\sigma_{k,d}} \leq \frac{\epsilon + 1 - \mu_{k,d}}{\sigma_{k,d}}\right) - \Pr\left(Z \leq \frac{\epsilon + 1 - \mu_{k,d}}{\sigma_{k,d}}\right)\right| = 0.$$

Property (a) of the Lemma (6.4) then gives part (ii).

6.8 Sketching asymptotics

We also wish to characterise the probability of obtaining an ϵ-subspace embedding for the Hadamard and Clarkson-Woodruff projections. Again let $\boldsymbol{A}$ be some large data matrix with singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\mathsf{T}$. The embedding probability of interest is

$$\Pr(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}) = \Pr(\sigma_{\max}\left(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}\right) \leq \epsilon).$$

Because of the discrete nature of the Hadamard and Clarkson-Woodruff projections it is cumbersome to express this probability in a meaningful way. As in Chapter 4 we again appeal to large n asymptotics to obtain an interpretable expression. Using the sketching central limit theorem (Theorem 4.3) we can argue that the Hadamard and Clarkson-Woodruff sketches have the same limiting embedding probability as the Gaussian projection. The sketching central limit theorem is restated here as Theorem 6.8. The necessary regularity conditions are described again in Assumption 1.

Assumption 1 Let the singular value decomposition of the $n \times d$ source dataset be given by $\boldsymbol{A}_{(n)} = \boldsymbol{U}_{(n)}\boldsymbol{D}_{(n)}\boldsymbol{V}_{(n)}^\mathsf{T}$. Let $\boldsymbol{u}_{(n)i}^\mathsf{T}$ give the ith row in $\boldsymbol{U}_{(n)}$ for $i = 1, \ldots, n$. Assume that the maximum leverage score tends to zero, that is

$$\lim_{n \to \infty} \max_{i=1,\ldots,r_n} \|\boldsymbol{u}_{(n)i}\|_2^2 = 0.$$

Theorem 6.8. *Consider a sequence of arbitrary $n \times d$ data matrices $\boldsymbol{A}_{(n)}$, where d is fixed. Let $\boldsymbol{A}_{(n)} = \boldsymbol{U}_{(n)}\boldsymbol{D}_{(n)}\boldsymbol{V}_{(n)}^\mathsf{T}$ represent the singular value decomposition of $\boldsymbol{A}_{(n)}$. Let $\boldsymbol{S}$ be a $k \times n$ Hadamard or Clarkson-Woodruff sketching matrix where k is also fixed. Let $\boldsymbol{u}_{(n)i}^\mathsf{T}$ represent the ith row of $\boldsymbol{U}_{(n)}$. Suppose that Assumption 1 on the maximum leverage score is satisfied. Then as n tends to infinity with k and d fixed,*

$$[\boldsymbol{S}\boldsymbol{A}_{(n)}]\boldsymbol{V}_{(n)}\boldsymbol{D}_{(n)}^{-1} \xrightarrow{d} \mathrm{MN}(\boldsymbol{0}, \boldsymbol{I}_k, \boldsymbol{I}_d/k).$$

Theorem 6.9 gives the asymptotic probability of obtaining an ϵ-subspace embedding for the Hadamard and Clarkson-Woodruff sketches.

Theorem 6.9. *Consider a sequence of arbitrary $n \times d$ data matrices $\boldsymbol{A}_{(n)}$, where each data matrix is of rank d, and d is fixed. Let $\boldsymbol{A}_{(n)} = \boldsymbol{U}_{(n)}\boldsymbol{D}_{(n)}\boldsymbol{V}_{(n)}^\mathsf{T}$ represent the singular value decomposition of $\boldsymbol{A}_{(n)}$. Let $\boldsymbol{S}$ be a $k \times n$ Hadamard or Clarkson-Woodruff sketching matrix where k is also fixed. Let $\boldsymbol{u}_{(n)i}^\mathsf{T}$ represent the ith row of $\boldsymbol{U}_{(n)}$. Assume that the maximum leverage score tends to zero, so*

$$\lim_{n \to \infty} \max_{i=1,\ldots,n} \|\boldsymbol{u}_{(n)i}\|_2^2 = 0.$$

Then as n tends to infinity with k and d fixed,

$$\lim_{n\to\infty} \Pr\left(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}_{(n)}\right) = \Pr\left(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W}) \leq \epsilon\right),$$

where $\boldsymbol{W} \sim \text{Wishart}(k, \boldsymbol{I}_d/k)$.

Proof: The sketching central limit theorem holds under Assumption 1. As we only need to consider the sequence of orthonormal matrices $\boldsymbol{U}_{(n)}$ to determine the embedding probability, we can use Theorem 6.8 with $\boldsymbol{D}_{(n)}$ and $\boldsymbol{V}_{(n)}$ set to the $d \times d$ identity matrix. As such we conclude that $\boldsymbol{S}\boldsymbol{U}_{(n)} \xrightarrow{d} MN(\boldsymbol{0}_{k\times d}, \boldsymbol{I}_k, \boldsymbol{I}_d/k)$. By the continuous mapping theorem it holds that for fixed d and k, asymptotically with n, $\boldsymbol{U}_{(n)}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}_{(n)} \xrightarrow{d} \text{Wishart}(k, \boldsymbol{I}_d/k)$. Another application of the continuous mapping theorem gives

$$\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}_{(n)}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}_{(n)}) \xrightarrow{d} \sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W}),$$

where $\boldsymbol{W} \sim \text{Wishart}(k, \boldsymbol{I}_d/k)$. We can use the continuous mapping theorem as the limiting Wishart matrix $\boldsymbol{W}$ has rank d with probability one. The maximum singular value function is continuous over the range where $\boldsymbol{W}$ has full rank (Bhatia, 1996). By the Portmanteau lemma it then holds that

$$\lim_{n\to\infty} \Pr\left(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}_{(n)}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}_{(n)}) \leq \epsilon\right) = \Pr\left(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W}) \leq \epsilon\right).$$

Now as

$$\begin{aligned}\lim_{n\to\infty} \Pr\left(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}_{(n)}\right) &= \lim_{n\to\infty} \Pr\left(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}_{(n)}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}_{(n)}) \leq \epsilon\right) \\ &= \Pr\left(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W}) \leq \epsilon\right),\end{aligned}$$

we have the final result □

The limiting embedding probability in Theorem 6.9 is the exact same embedding probability as for the Gaussian sketch (6.7). As n grows, we expect the Hadamard and Clarkson-Woodruff sketches to be as effective as the Gaussian sketch for generating ϵ-subspace embeddings. This is significant as the Hadamard and Clarkson-Woodruff sketches are dramatically faster than the Gaussian projection (see Table 6.1). The Clarkson-Woodruff sketch is $O(k)$ times faster than the Gaussian projection, which can be a very large factor in practice.

We can again use the Tracy-Widom distribution to approximate the limiting embedding probability $\Pr\left(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W}) \leq \epsilon\right)$ for the Clarkson-Woodruff and Hadamard sketches. We are unable to establish a formal limit theorem in terms of the Tracy-Widom distribution

$$\lim_{n,d,k\to\infty} \Pr(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}_{(n)}) = \Pr\left(Z \leq \frac{\epsilon + 1 - \mu_{k,d}}{\sigma_{k,d}}\right),$$

where $Z \sim F_1$ as per Theorem 6.7. This is because the sketching central limit theorem (Theorem 6.8) has k and d fixed, with only n being taken to infinity. Central limit theorems under expanding dimension are more challenging (Portnoy, 1986), and this is left as future work. It is possible that Assumption 1 on the leverage scores will remain sufficient in the expanding dimension scenario. The following reasoning was also used by Huber (1973) in the analysis of high dimensional regression models. For any d, the maximum leverage score must be greater than the average leverage score. We thus have

$$\begin{aligned}\max_{i=1,\ldots,n} \|\boldsymbol{u}_{(n)i}\|_2^2 &\geq \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{u}_{(n)i}\|_2^2 \\ &= \frac{d}{n}.\end{aligned}$$

If we maintain that Assumption 1 holds on the leverage scores as $n, d, k \to \infty$ we necessarily imply that $d/n \to 0$. This is intuitively reasonable. If we simultaneously require $d/k \to \alpha$ where $\alpha \in (0, 1]$ we must

also have that $k/n \to 0$. This is also a sensible requirement. The proofs in Chapter 5 do not appear to have any critical weaknesses that prevent an extension to the expanding dimension scenario.

The lack of a central limit theorem in the expanding dimension scenario is not a critical gap, as the key result is that the Hadamard and Clarkson-Woodruff sketches behave like the Gaussian projection for large n, with k and d fixed. If the Tracy-Widom approximation in Theorem 6.7 is good for finite k and d with the Gaussian sketch, it should hold well for the Hadamard and Clarkson-Woodruff projections for n sufficiently large.

6.9 Data application

We test the theory on a large genetic dataset. The covariate data consists of genotypes at $p = 1032$ genetic variants on $n = 407,779$ subjects. The genetic variants are in the Protein Kinase C Epsilon (PRKCE) gene. Variants were filtered to have mean allele frequency of greater than one percent. The response variable is haemoglobin concentration adjusted for age/sex and technical covariates. We also consider a subset of this dataset with $p = 130$ representative markers identified by hierarchical clustering.

We first assess the accuracy of Theorem 6.7 and 6.9 by comparing the theoretical embedding probability to the observed embedding probability. We then check the accuracy of the posterior approximation as a function of ϵ in a simulation study. We compare the Gaussian, Hadamard and Clarkson-Woodruff sketches. We also compare the data oblivious sketches to simple random sampling with replacement. The simple random sampling of observations is also referred to as the Uniform projection in the sketching literature. As mentioned in Chapter 4, although the uniform projection is simple to understand and implement, it is difficult to establish strong error bounds on its performance.

6.9.1 Embedding probabilities

The dataset is of moderate size, so it is feasible to take the singular value decomposition of the full $n \times d$ dataset $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\mathsf{T}$. Given the singular value decomposition we can run an oracle procedure to determine the empirical embedding probability. Suppose we take B sketches. Let $\boldsymbol{S}^{[1]}, \ldots, \boldsymbol{S}^{[B]}$ denote the B sketching matrices. Define $\epsilon^{[b]}$ as

$$\epsilon^{[b]} = \sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^{[b]\mathsf{T}}\boldsymbol{S}^{[b]}\boldsymbol{U}),$$

for $b = 1, \ldots, B$. The value $\epsilon^{[b]}$ measures the quality of the sketching matrix $\boldsymbol{S}^{[b]}$. The estimated embedding probability is then

$$\begin{aligned}\widehat{\Pr}(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}) &= \widehat{\Pr}(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}) \leq \epsilon) \\ &= \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}(\epsilon^{[b]} \leq \epsilon).\end{aligned}$$

Figure 6.6 shows the empirical and theoretical embedding probabilities for the representative PRKCE dataset ($n = 407,779, d = 132$) for each type of sketch. We took one hundred sketches at $k = 20 \times d$. The observed and theoretical curves match well for the Gaussian, Hadamard and Clarkson-Woodruff projection. The uniform projection performs worse than the other data-oblivious random projections. The distribution of ϵ is shifted to the right compared to the other projections. Larger values of ϵ indicate weaker approximation bounds. The uniform projection does not satisfy a central limit theorem for fixed k, so we do not necessarily expect the Tracy-Widom law to give a good approximation for the uniform projection. Table 6.2 reports the average sketching time for each projection. The Clarkson-Woodruff projection is much faster than the Gaussian projection, but has the same embedding probability, as predicted by Theorem 6.9.

Figure 6.7 shows the empirical and theoretical embedding probabilities for the full PRKCE dataset ($n = 407,779, d = 1034$). We took one hundred sketches at $k = 20 \times d$. The x-axis is different in

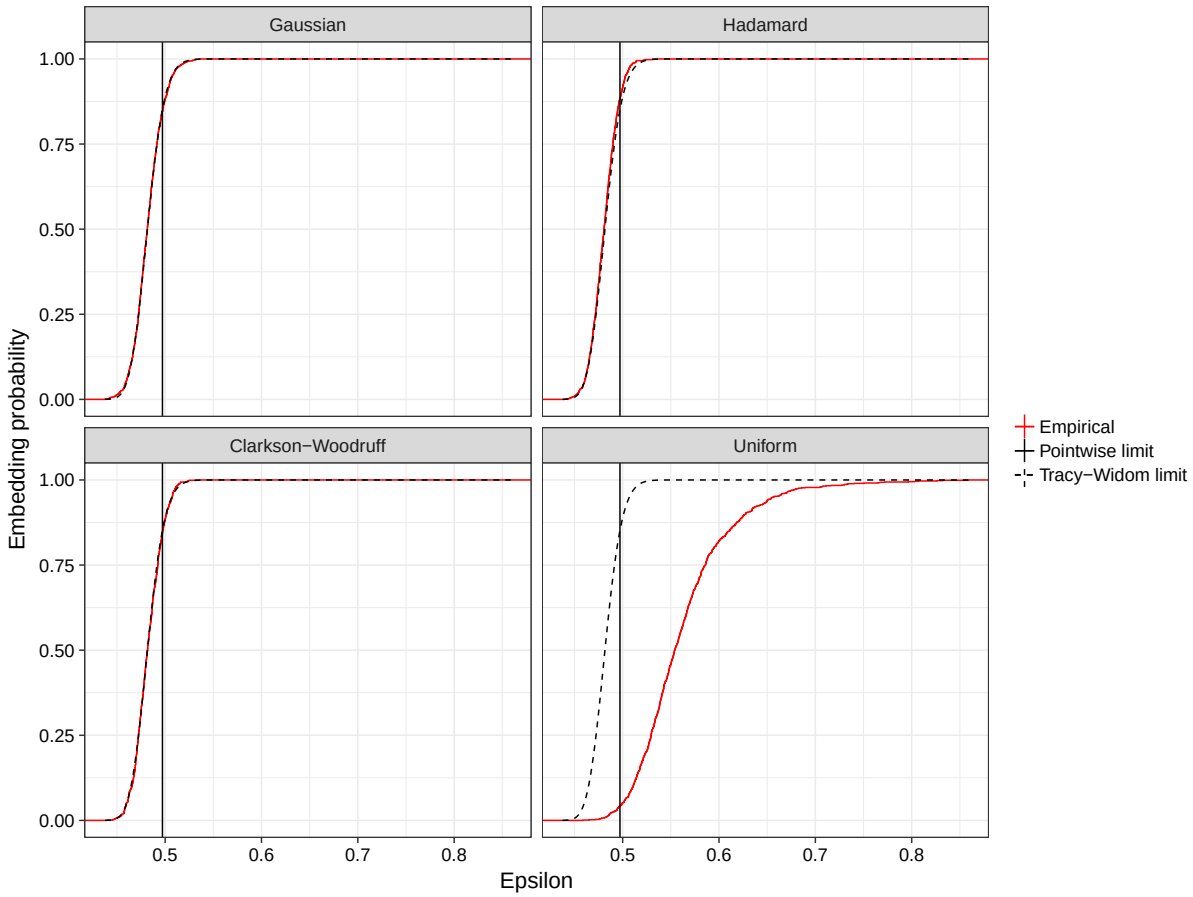


Figure 6.6: Empirical and theoretical embedding probabilities for the representative PRKCE dataset. Hierarchical clustering was used to select a subset of genetic variants from the full PRKCE dataset. We want the sketching matrix $\boldsymbol{S}$ to be an ϵ-subspace embedding for the source dataset with small ϵ with high probability. The y-axis gives the proportion of times we attain an ϵ-subspace embedding. The x-axis gives the distortion factor ϵ (recall Definition 6.1). In this simulation $n = 407,779, d = 132, k = 20 \times d$. One thousand sketches were generated for each type of sketch. The vertical line gives the asymptotic pointwise limit. The Tracy-Widom approximation is very accurate for the Gaussian, Hadamard and Clarkson-Woodruff sketches.

Projection	Gaussian	Hadamard	Clarkson-Woodruff	Uniform
Time (seconds)	769	17.6	1.34	0.03

Table 6.2: Mean sketching time for the representative PRKCE dataset. Results for the Hadamard, Clarkson-Woodruff and Uniform projections are over one hundred sketches. Results for the Gaussian sketch are over ten sketches.The Gaussian sketch is considerably slower than the Hadamard and Clarkson-Woodruff sketches as is expected from Table 6.1.

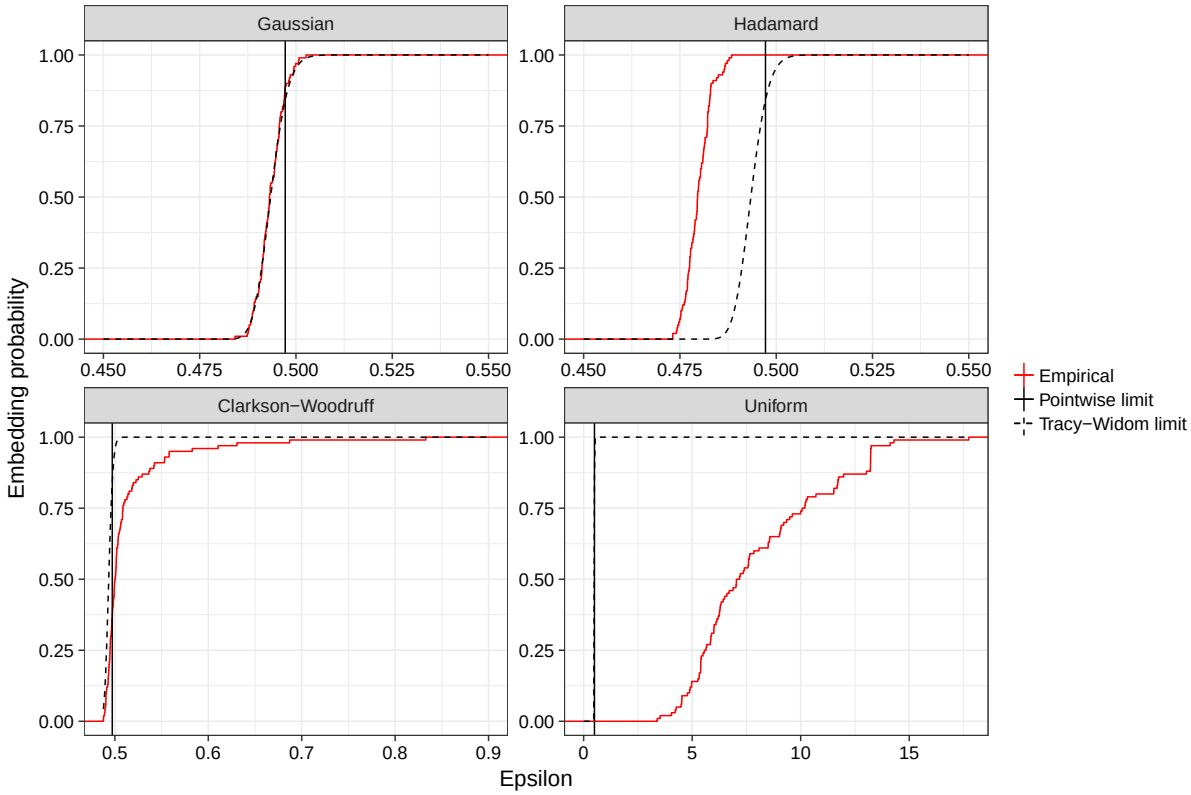


Figure 6.7: Empirical and theoretical embedding probabilities for the full PRKCE genetic dataset ($n = 407,779, d = 1034, k = 20 \times d$). One hundred sketches were generated for each type of sketch. We want the sketching matrix $\boldsymbol{S}$ to be an ϵ-subspace embedding for the source dataset with small ϵ with high probability. The y-axis gives the proportion of times we attain an ϵ-subspace embedding. The x-axis gives the distortion factor ϵ (recall Definition 6.1). Different x-scales are used in each plot. The Tracy-Widom approximation is very accurate for the Gaussian sketch. There is moderate deviation from the asymptotic approximation for the Hadamard sketch and larger deviation for the Clarkson-Woodruff sketch.

each plot. Overall, the empirical and theoretical curves do not match as well compared to Figure 6.6. The approximation for the Gaussian projection is very accurate. Interestingly, the Hadamard projection slightly outperforms the Gaussian projection. The dashed line in the Hadamard panel is shifted slightly to the left compared to the theoretical curve, representing better than expected performance. The Clarkson-Woodruff projection has a much longer right tail then is predicted by the Tracy-Widom law. The Uniform projection does very poorly in this simulation compared to the the other data oblivious projections. Even though the distribution of ϵ under Clarkson-Woodruff projection has a longer right tail than the theoretical, the distortion factors are always below 1. The median for the Uniform projection is above 5. The Hadamard and Clarkson-Woodruff projections to be more stable in this example. Data oblivious projections are designed to be robust, and it appears there are features in this dataset that cause major problems for uniform subsampling that are less of an issue for the data oblivious sketches. The extra computational cost of the Hadamard and Clarkson-Woodruff projections come with some demonstrable benefits.

Table 6.3 reports the average sketching time for each projection. The Gaussian sketch was not timed in this example as we simulated directly from the matrix normal distribution of the sketched data $\widetilde{\boldsymbol{A}} \sim MN(\boldsymbol{0}, \boldsymbol{I}_k, \boldsymbol{A}^\mathsf{T}\boldsymbol{A}/k)$ rather than computing the matrix product $\boldsymbol{SA}$ directly. This was so that the simulation could be run in a reasonable amount of time. The Clarkson-Woodruff sktch is again faster than the Hadamard sketch as is expected.

We can better understand the deviation from the Tracy-Widom limit in Figure 6.7 by looking at the density of $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})$. Figure 6.8 compares the empirical density to the Tracy-Widom approximation given in Theorem 6.7. The empirical distribution under the Hadamard sketch is shifted to

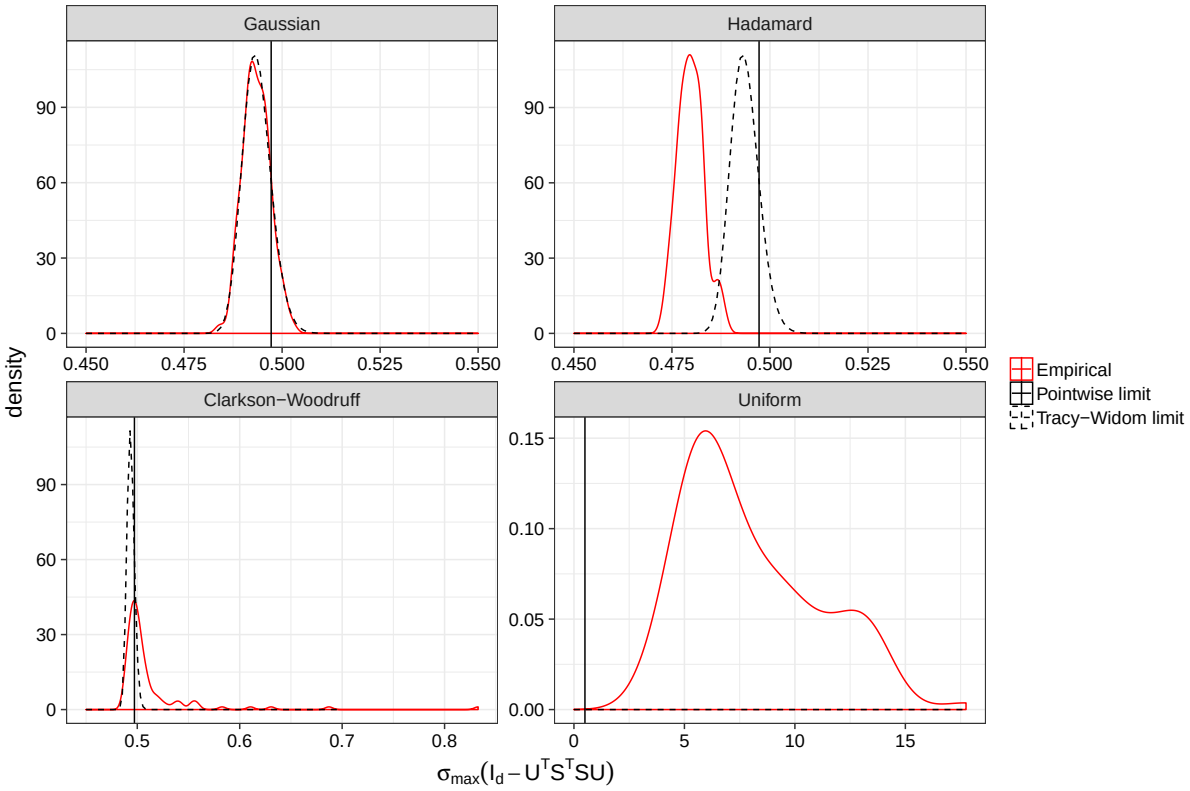


Figure 6.8: Empirical and theoretical density of $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})$ for the full PRKCE genetic dataset. We want the sketching matrix $\boldsymbol{S}$ to be an ϵ-subspace embedding for the source dataset with small ϵ with high probability. The sketching matrix $\boldsymbol{S}$ is an ϵ-subspace embedding for the source dataset if and only if $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}) \leq \epsilon$. One hundred sketches were generated for each type of sketch. The vertical line gives the asymptotic pointwise limit. Different x-scales are used in each plot. The Tracy-Widom approximation is very accurate for the Gaussian sketch. There is moderate deviation from the asymptotic approximation for the Hadamard sketch and larger deviation for the Clarkson-Woodruff sketch.

the left compared to the asymptotic result, and the empirical distribution under the Clarkson-Woodruff sketch has more right skew.

The deviation from the Tracy-Widom limit in Figure 6.8 could be because the finite sample approximation is poor. Theorem 6.9 suggests that the Hadamard and Clarkson-Woodruff projections behave like the Gaussian sketch for n sufficiently large. To test this we bootstrapped the full PRKCE dataset to be ten times its original size. The bootstrapped PRKCE dataset has $n = 4,077,790, d = 1034$. We took one thousand sketches of size $k = 20 \times d$ using the Clarkson-Woodruff projection and ran the oracle procedure of computing $\epsilon^{[b]} = \sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^{[b]\mathsf{T}}\boldsymbol{S}^{[b]}\boldsymbol{U})$ for each sketch. Figure 6.9 compares the distribution of $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})$ using Clarkson-Woodruff projection on the original dataset and on the large bootstrapped dataset. As n increases we expect the quality of the Tracy-Widom approximation to improve. Panel (a) compares the theoretical to the simulation results on the original dataset. The Clarkson-Woodruff projection shows greater variance than expected. Panel (b) compares the theoretical to the simulation results on the bootstrapped dataset. In (b) we see very good agreement between the empirical distribution and the theoretical distribution. It seems that for this dataset $n \approx 400,000$ is not big enough for the large sample asymptotics to kick in. At $n \approx 4$ million we see the expected asymptotic behaviour. The results in Figure 6.9 are significant as they reflect the culmination of the asymptotic analysis in Chapter 5 and the asymptotic results in this chapter. The sketching central limit theorem ties the large n behaviour of the Clarkson-Woodruff projection to that of the Gaussian projection. Theorem 6.7 links the large d and k behaviour of the Gaussian projection to the Tracy-Widom law. The finite sample analysis of the Clarkson-Woodruff projection is very difficult. By chaining together asymptotic results we can obtain asymptotic predictions on its performance. As with any asymptotic analysis there

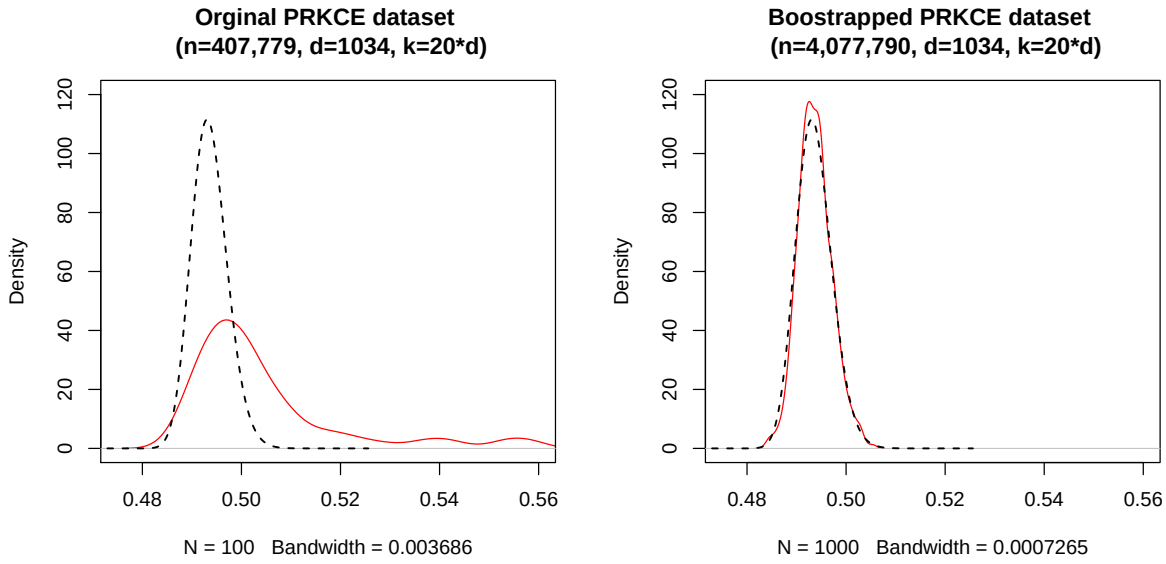


Figure 6.9: Comparison of theoretical and empirical density of $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\top \boldsymbol{S}^\top \boldsymbol{S}\boldsymbol{U})$ on the full PRKCE dataset $(n = 407,779)$ and the bootstrapped PRKCE dataset $(n = 4,077,790)$. We want the sketching matrix $\boldsymbol{S}$ to be an ϵ-subspace embedding for the source dataset with small ϵ with high probability. The sketching matrix $\boldsymbol{S}$ is an ϵ-subspace embedding for the source dataset if and only if $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\top \boldsymbol{S}^\top \boldsymbol{S}\boldsymbol{U}) \leq \epsilon$. The dashed black line gives the Tracy-Widom limit and the solid red line shows a kernel density estimate. The sketch size was set at $k = 20 \times d$, where $d = 1034$. We used a Clarkson-Woodruff sketch. The accuracy of the Tracy-Widom approximation improves as n increases as is expected from the asymptotic theory (Theorem 6.9).

Projection	Gaussian	Hadamard	Clarkson-Woodruff	Uniform
Time (seconds)	-	156	21	2.8

Table 6.3: Mean sketching time (seconds) for the full PRKCE dataset $(n = 407,779, d = 1034)$. This dataset is too large to use Gaussian sketch directly. The Clarkson-Woodruff sketch is faster than the Hadamard sketch as is expected from Table 6.1.

is there is the question of how large n has to be before the approximation is reasonable. Further study on the rate of convergence and possible finite sample bounds on the error in the normal approximation would be very useful and interesting research directions. We have used asymptotic theory to assess the embedding probability (6.1). The results in Figure 6.9 show the asymptotic approximations are reasonable in a realistic data setting. Given the rapidly decreasing cost of genotyping, it is reasonable to anticipate the n and d in panel (b) to be commonplace in multivariable genotype-phenotype studies in the near future. To integrate sketching methods in the analysis pipeline it is necessary to understand the level of error introduced by using the random projection. The asymptotic results are useful as they provide important guidance on the quality of the randomised data compression step.

Table 6.3 reports the average sketching time for each projection on the full dataset. This is the time required to compute the random sketched dataset $\widetilde{\boldsymbol{A}} = \boldsymbol{S}\boldsymbol{A}$. The Clarkson-Woodruff sketch is faster than the Hadamard sketch as is expected from Table 6.1.

6.9.2 Posterior approximation

We also compare the accuracy of the sketched posterior distribution at different sketch sizes. Sketching is a tool for approximate computation. As the sketch size k increases the accuracy of the approximate calculation increases. As discussed in section 6.4 this behaviour can be formalised using ϵ-subspace embeddings. As k increases we $\epsilon \xrightarrow{p} 0$ and the sketched posterior distribution will approach the target posterior distribution. Here we examine the sensitivity of the results to k and ϵ in a practical setting.

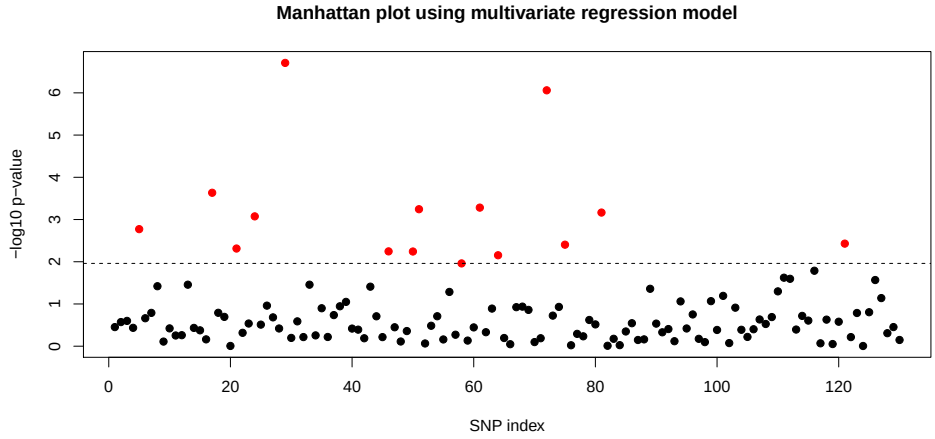


Figure 6.10: Manhattan plot using the representative PRKCE dataset. The y-axis gives the minus log 10 p-value for each predictor when testing a null hypothesis of zero. We fit the saturated model using $p = 130$ markers and performed a t-test on each coefficient associated with a particular genetic marker.

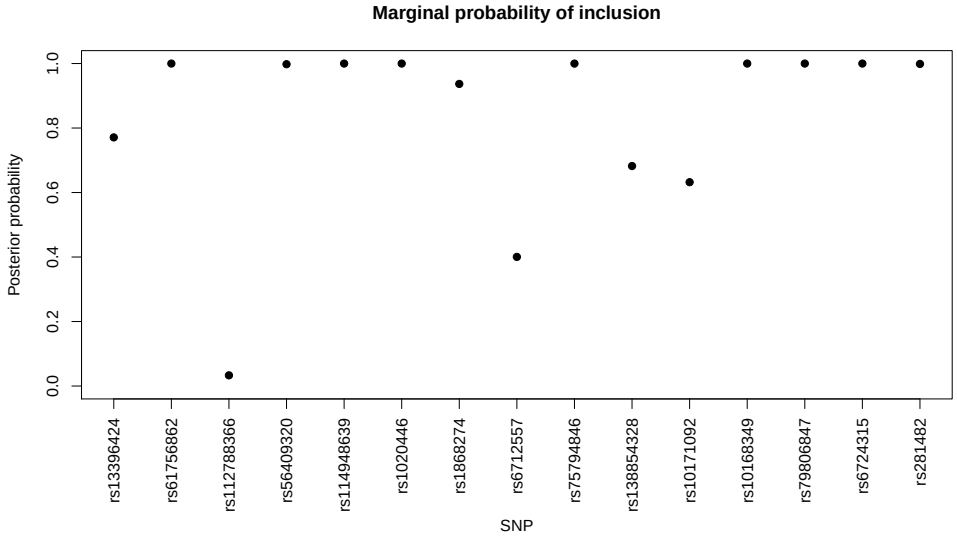


Figure 6.11: Marginal inclusion probabilities from an analysis using the top 15 SNPs. These results were obtained by enumerating over all models using the full dataset. There are some variants with low support for inclusion, some variants with moderate support for inclusion and some variants with high support for inclusion. The range of target marginal inclusion probabilities makes this a useful benchmark dataset for sketching.

To isolate the error introduced by the sketch we limit our analysis to a dataset with $p = 15$ genetic variants so that the posterior can be computed by enumeration. This is so there is not Monte Carlo error in the target posterior distribution. We took the top 15 variants identified by fitting a multivariable regression model using the representative PRKCE dataset. Figure 6.10 shows a Manhattan plot of the $p = 130$ genetic markers in the dataset. We take the 15 markers with the smallest p-values. The dashed line in the plot shows the cutoff used. The design matrix $\boldsymbol{X}$ was then set to only contain the top 15 genetic variants. From here we computed the target posterior over models $\boldsymbol{\gamma}$. There are $2^{16} = 65536$ candidate models as we allow for an intercept. Figure 6.11 shows the marginal posterior probability of inclusion for each of the top 15 variants in the restricted analysis. We fixed the error variance at $\sigma^2 = n/(n - 131)\widehat{\sigma}^2$ where $\widehat{\sigma}^2$ is the maximum likelihood estimate of the residual variance from fitting the saturated model with $p = 130$ genetic predictors. We then computed the integrated likelihood for all models using (6.2). By normalising we obtained the the exact posterior distribution on the full dataset. We benchmark the sketched posterior against the true posterior distribution.

We then took $B = 100$ hundred sketches at different sizes k using the Clarkson-Woodruff projection. Let $\widetilde{\boldsymbol{A}}^{[b]}$ be the bth sketched dataset for $b = 1, \ldots, B$. For $b = 1, \ldots, B$ we then computed the sketched sufficient statistics $\widetilde{\boldsymbol{A}}^{[b]\mathsf{T}}\widetilde{\boldsymbol{A}}^{[b]}$. We then computed the approximate integrated likelihood for each model using (6.5). Formally, for $b = 1, \ldots, B$ and all models $\boldsymbol{\gamma}$ we calculated

$$\widetilde{p}^{[b]}(\boldsymbol{y}|\boldsymbol{\gamma}, g, \sigma^2) = (1+g)^{-p_\gamma/2} \exp\left[-\frac{1}{2\sigma^2}\left(\frac{g}{g+1}(RSS_S^\gamma)^{[b]} + \frac{1}{g+1}\boldsymbol{y}^\mathsf{T}\boldsymbol{y}\right)\right]. \tag{6.20}$$

Using the approximate integrated likelihoods we obtain a sketched posterior over models $\widetilde{p}^{[b]}(\boldsymbol{\gamma}|\boldsymbol{y}, g, \sigma^2)$ for $b = 1, \ldots, B$. We computed the sketched marginal probabilities of inclusion for $b = 1, \ldots, B$. We compare $\widetilde{p}^{[b]}(\boldsymbol{\gamma}|\boldsymbol{y}, g, \sigma^2)$ for $b = 1, \ldots, B$ to the target posterior $\widetilde{p}(\boldsymbol{\gamma}|\boldsymbol{y}, g, \sigma^2)$ using the marginal probabilities of inclusion. Figure 6.12 shows boxplots of the sketched posterior results over the $B = 100$ sketches. Each panel shows the results for the marginal inclusion probability for a particular genetic variant. The red dashed line gives the target marginal inclusion probability. At a high-level, as the sketch size increases, the quality of the approximate posterior should increase. Here quality refers to the difference compared to the exact posterior distribution computed on the full dataset $\widetilde{p}(\boldsymbol{\gamma}|\boldsymbol{y}, g, \sigma^2)$. The gold-standard results are given by the red-dashed line in each panel. We would like to see boxplots tightly concentrated around the red dashed line. We do not see this. The panel in row 1 column 3 is worth studying in detail. The target marginal inclusion probability is below 0.05. The sketched approximations return marginal inclusion probabilities close to one for k up to ten thousand. The quality of the approximate calculation is very poor for this SNP. The panel in row 3 column 2 is also worth examining. The target marginal inclusion probability is around 0.45. The sketched marginal inclusion probabilities are biased upwards for k up to ten thousand. The sketched inclusion probabilities remain highly variable at k equal to one hundred thousand.

The variance of the sketched approximation is very high and it is hard to discern any clear trends due tot the use of boxplots. Figures 6.13 and 6.14 show histograms of the approximate marginal inclusion probabilities for selected SNPs. Looking at the first column in Figure 6.13 we can see a large positive bias at $k = 100$. The sketched inclusion probabilities are bunched around one when the target inclusion probabilities are below 0.5. The results in Figure 6.14 show some interesting patterns. At $k = 100$ thousand we see that the sketched posterior inclusion probabilities have a U shaped distribution with modes at zero and one. We would like to see a mode around the red-dashed line. The noise introduced by the sketch has a strong and perhaps unpredictable effect on the marginal includion probabilities.

The Gaussian sketch does not require that k be smaller than n. We repeated the analysis with very large sketch sizes to identify the necessary k to give a tolerable posterior approximation. We would also like to see if smoother behaviour emerges at larger sketch sizes. We simulated directly from the distribution of the sketched sufficient statistics $\widetilde{\boldsymbol{A}}^\mathsf{T}\widetilde{\boldsymbol{A}} \sim \text{Wishart}(k, \boldsymbol{A}^\mathsf{T}\boldsymbol{A}/k)$ to bypass the huge computational cost of generating comically large sketches. Figure 6.15 shows boxplots of the approximate marginal inclusion probabilities. Figures 6.16 and 6.17 show histograms of the approximate marginal inclusion probabilities for selected SNPs. These results show the general dynamic that we expect. As k increases we see concentration around the target marginal inclusion probabilities (red dashed line). In the fourth column in 6.17 we see a unimodal distribution centered around the target marginal inclusion probability rather than the U shaped distribution that was noted in the fourth column of Figure 6.14.

Even at $k = 10$ million there is significant variance around the target inclusion probability. At $k = 1$ billion the error looks acceptable. The posterior distribution over models is a complicated nonlinear function of the sufficient statistics. Small changes in the approximated sufficient statistics can lead to a large change in the approximated posterior distribution. It is hard to determine apriori what level of deviation we expect in the approximate posterior given a particular sketch size.

As before, we can run an oracle procedure to measure the quality of sketch at each k. Let the singular value decomposition of the source dataset again be given by $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\mathsf{T}$. Define

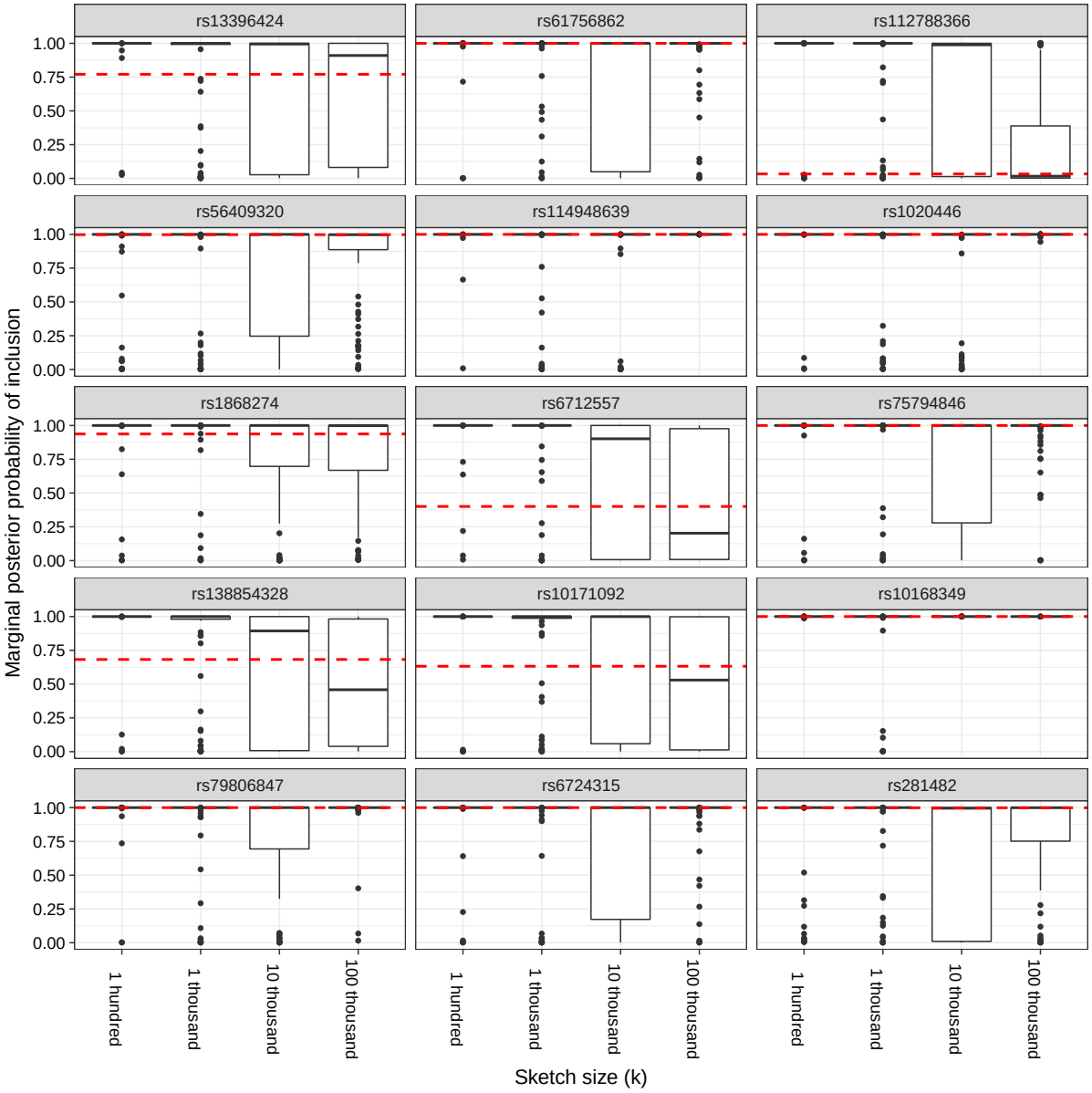


Figure 6.12: Boxplots of sketched marginal posterior probabilities of inclusion. The Clarkson-Woodruff projection was used at different sketch sizes k. One hundred sketches were taken at each value of k. The target marginal probability of inclusion for each SNP is plotted as red dashed line. The sketched posterior approximations should approach the target values as k increases. The variability in the sketched approximation differs over SNPs. In general the sketched estimates lie above the target values (the red dashed line). It is hard to see a clear relationship between the variance of the sketched approximations and k in this plot.

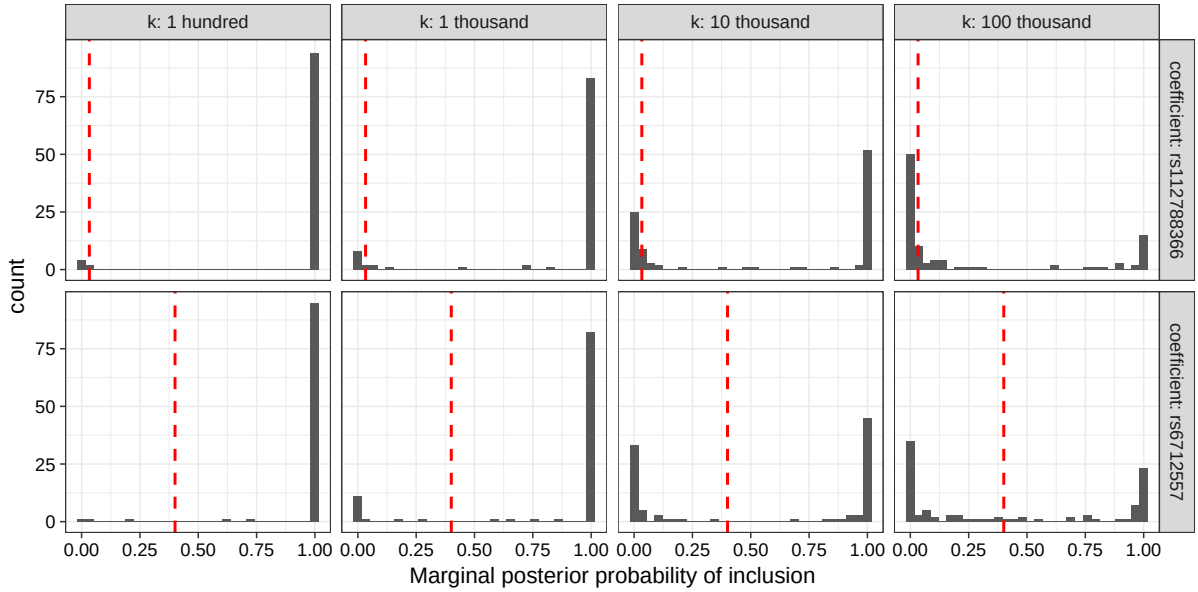


Figure 6.13: Histograms of sketched marginal posterior probabilities of inclusion for SNPs where the target marginal probability is less than 0.5. Results were obtained using the Clarkson-Woodruff projection at different sketch sizes k. The target marginal probability of inclusion for each SNP is plotted as red dashed line. The sketched posterior approximations should approach the target values as k increases. The sketched marginal inclusion probabilities are concentrated around zero and one in each panel. It is hard to see a clear relationship between the variance of the sketched approximations and k in this plot.

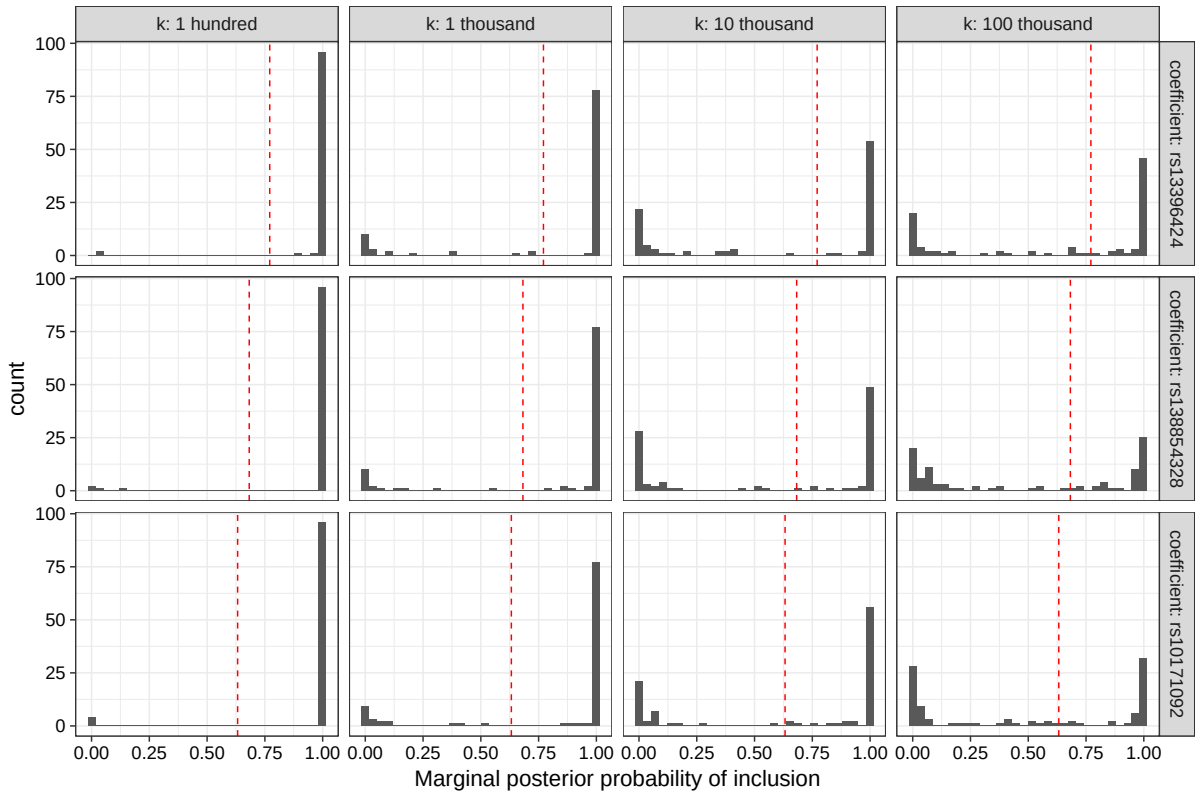


Figure 6.14: Histograms of sketched marginal posterior probabilities of inclusion for SNPs where the target marginal probability is between 0.5 and 0.9. Results were obtained using the Clarkson-Woodruff projection at different sketch sizes k. The target marginal probability of inclusion for each SNP is plotted as red dashed line. The sketched posterior approximations should approach the target values as k increases. The sketched marginal inclusion probabilities are concentrated around zero and one in each panel.It is hard to see a clear relationship between the variance of the sketched approximations and k in this plot.

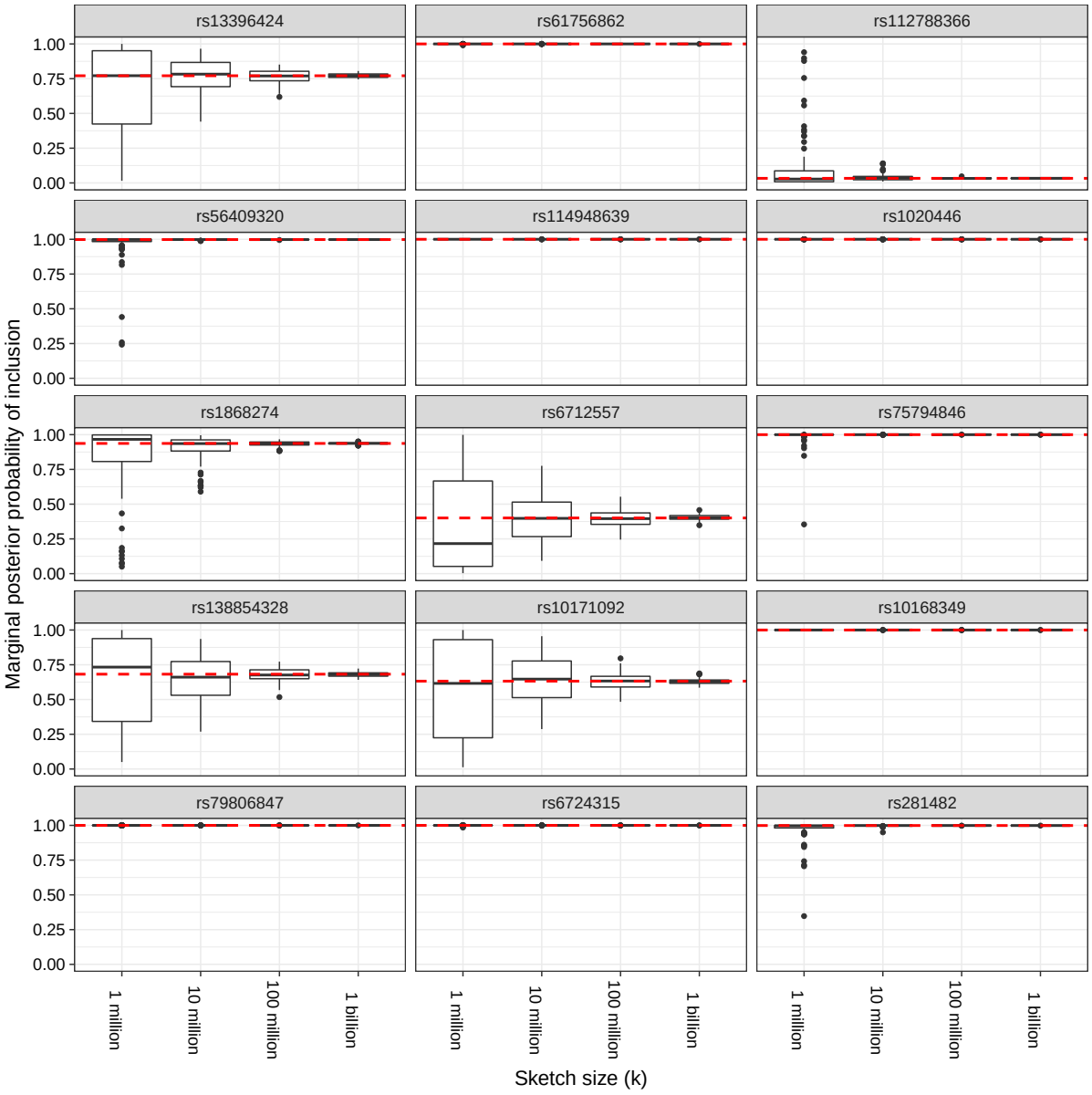


Figure 6.15: Boxplots of sketched marginal posterior probabilities of inclusion. The target marginal probability of inclusion for each SNP is plotted as red dashed line. The Gaussian sketch was used at different sketch sizes k. The sketched posterior approximations should approach the target values as k increases. One hundred sketches were taken at each value of k. As the sketch size k increases the approximate marginal inclusion probabilities concentrate around the target values.

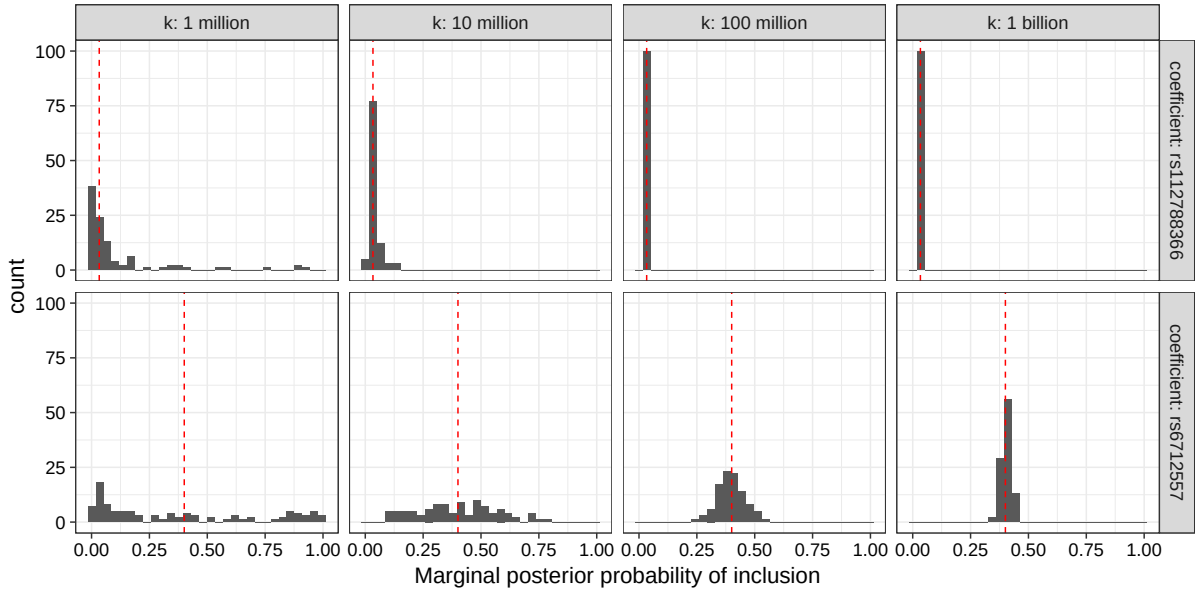


Figure 6.16: Histograms of sketched marginal posterior probabilities of inclusion for SNPs where the target marginal probability is less than 0.5. Results were obtained using the Gaussian projection at different sketch sizes k. The target marginal probability of inclusion is plotted as a dashed line. The sketched posterior approximations should approach the target values as k increases. As the sketch size k increases the approximated marginal inclusion probabilities concentrate around the target values.

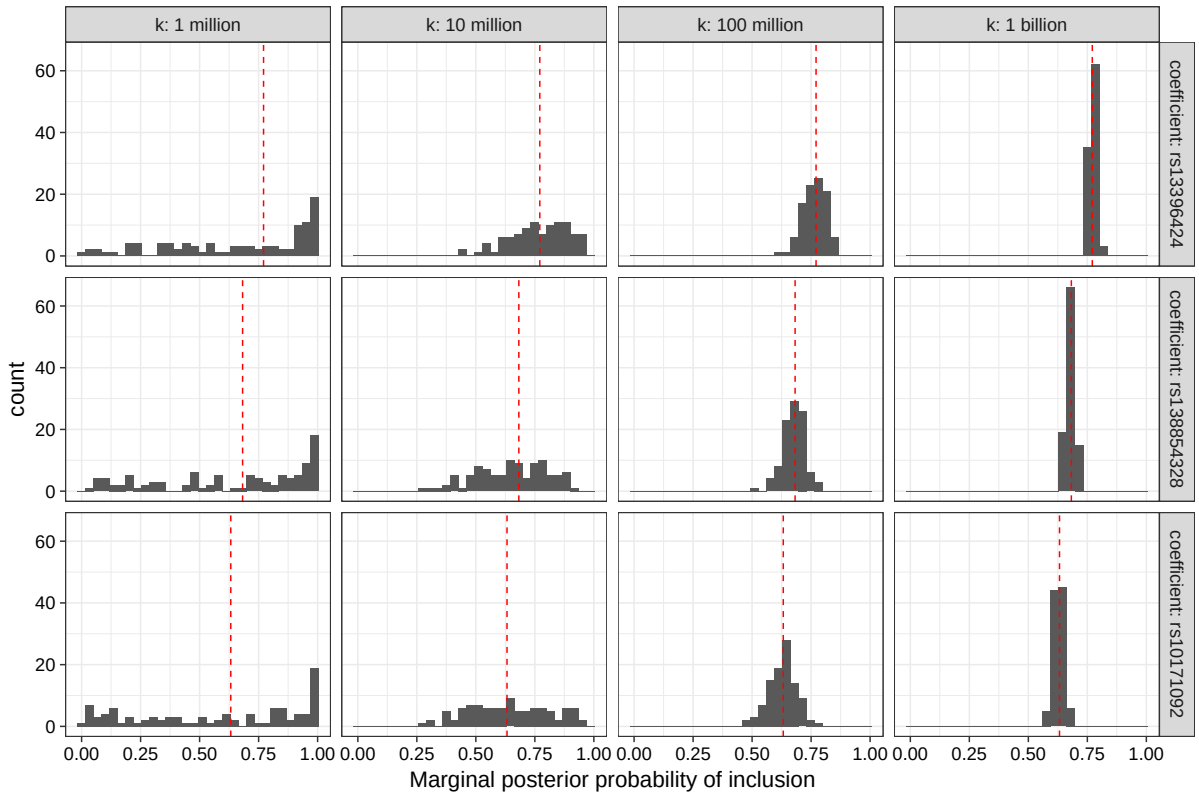


Figure 6.17: Histograms of sketched marginal posterior probabilities of inclusion for SNPs where the target marginal probability is between 0.5 and 0.9. Results were obtained using the Gaussian projection at different sketch sizes k. The target marginal probability of inclusion is plotted as a dashed line. The sketched posterior approximations should approach the target values as k increases. As the sketch size k increases the approximated marginal inclusion probabilities concentrate around the target values.

$$\bar{\epsilon} = \mathbb{E}_{\boldsymbol{S}}[\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})].$$

The value $\bar{\epsilon}$ gives the expected distortion factor over the random sketch. This can be compared to the values predicted by Theorem 6.5 and Theorem 6.7. Using the Tracy-Widom approximation we expect that

$$\bar{\epsilon} \approx \sigma_{k,d}\mathbb{E}[\boldsymbol{Z}] + \mu_{k,d} - 1,$$

where $\mu_{k,d}$ and $\sigma_{k,d}$ are the centring and scaling constants in Theorem 6.6, and $\mathbb{E}[\boldsymbol{Z}]$ is the mean of a Tracy-Widom random variable, $\mathbb{E}[\boldsymbol{Z}] \approx -1.21$. Using the more crude pointwise approximation in Theorem 6.5 we have $\bar{\epsilon} \approx (1 + \sqrt{d/k})^2 - 1$. A Monte Carlo estimate of $\bar{\epsilon}$ is obtained by taking the sample average of $\epsilon^{[1]}, \ldots, \epsilon^{[B]}$. Table 6.4 compares the sample average to the asymptotic predictions for the Clarkson-Woodruff projection. There is a good correspondence between the theoretical and observed values. Table 6.5 compares the sample average to the asymptotic predictions for the Gaussian sketch. The Tracy-Widom approximation from Theorem 6.6 gives more accurate predictions than the pointwise approximation in Theorem 6.5. We are able to estimate the expected distortion factor $\bar{\epsilon}$ for a particular sketch size k using the asymptotic theory. The sketches are behaving as expected in terms of the distortion factor ϵ. It seems that very small ϵ is required to give a tolerable posterior approximation. We can determine ahead of time what ϵ is expected for a given sketch size k. This can be used to set expectations appropriately when interpreting the the sketched results.

Sketch size (k)	one hundred	one thousand	ten thousand	one hundred thousand
Monte Carlo estimate of $\bar{\epsilon}$	0.857	0.244	0.079	0.024
Tracy-Widom expectation	0.851	0.245	0.075	0.023
Pointwise expectation	0.995	0.278	0.084	0.026

Table 6.4: Comparison of Monte Carlo and asymptotic estimates of $\bar{\epsilon} = \mathbb{E}_{\boldsymbol{S}}[\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})]$ for the Clarkson-Woodruff sketch. The quality of a sketch can be summarised by $\bar{\epsilon}$. The source dataset is the top 15 SNP dataset described in section 6.9.2 ($n = 407,779, d = 17$). The Tracy-Widom expectations (Theorem 6.7) match the Monte Carlo estimates more closely than the pointwise limit (Theorem 6.5). We are able to estimate the expected distortion factor $\bar{\epsilon}$ for a particular sketch size k using the asymptotic theory.

Sketch size (k)	one million	ten million	one hundred million	one billion
Monte Carlo estimate of $\bar{\epsilon}$	0.00771	0.00247	0.00077	0.00025
Tracy-Widom expectation	0.00738	0.00233	0.00074	0.00023
Pointwise expectation	0.00826	0.00261	0.00082	0.00026

Table 6.5: Comparison of Monte Carlo and asymptotic estimates of $\bar{\epsilon} = \mathbb{E}_{\boldsymbol{S}}[\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})]$ for the Gaussian sketch. The quality of a sketch can be summarised by $\bar{\epsilon}$. The source dataset is the top 15 SNP dataset described in section 6.9.2 ($n = 407,779, d = 17$). The Tracy-Widom expectations (Theorem 6.7) match the Monte Carlo estimates more closely than the pointwise limit (Theorem 6.5). We are able to estimate the expected distortion factor $\bar{\epsilon}$ for a particular sketch size k using the asymptotic theory.

6.10 Conclusion

We have investigated the use of sketching for approximate Bayesian subset selection. Approximation bounds are developed using ϵ-subspace embeddings. Data oblivious random projections are designed to output an ϵ-subspace embedding of the source dataset with high probability. We obtained asymptotic

expressions for the embedding probability for the Gaussian, Hadamard and Clarkson-Woodruff sketch in terms of the Tracy-Widom law. Simulations showed that the asymptotic expression is accurate on large datasets. The asymptotic theory can be used to estimate the required sketch size k needed to obtain an ϵ-subspace embedding with probability at least $(1 - \delta)$. Determination of the embedding probability $\Pr(\boldsymbol{S}$ is an ϵ-subspace embedding for $\boldsymbol{A})$ is a largely open question, and the asymptotic equivalence of the Gaussian, Hadamard and Clarkson-Woodruff sketched is of theoretical and practical interest. The universality of the Tracy-Widom law is a topic of current research (Bao et al., 2015) and it is also of interest to determine the full set of data oblivious random projections that satisfy the Tracy-Widom limiting embedding probability.

We do not think that asymptotic analysis can supplant the finite sample results that drive much of the existing work on randomised algorithms. We feel that they are a useful complement that can provide answers to important questions that are nigh unresolvable in a finite sample framework. A combination of asymptotic and non-asymptotic results could be useful for the development of more advanced algorithms.

We also investigated the range of ϵ needed to obtain a tolerable posterior approximation. Sketching has been proposed for approximating the posterior distribution over coefficients in the fixed model setting (Geppert et al., 2017). We found that ϵ needs to be much smaller than in the fixed model setting to maintain the important features of the target posterior over models. It appears that approximation of the integrated likelihood is a more difficult task than preservation of the normal likelihood conditional on a model.

7

Conclusion

As the thesis winds down, we can ease into the customary retrospection by cycling back to the introduction and reconsidering Box's loop (Figure 7.1). Box's loop breaks down a generic data analysis problem into a series of connected sub-problems involving model formulation, fitting, checking and revising. This staged workflow is described by Blei (2014) as build, compute, critique and repeat. The modularisation helped us to pin down the central research question of interest. The emergence of Big Data has led to the compute step becoming a bottleneck in practical applications. As datasets grow larger, the computational budget does not stretch as far and we are forced to adapt. New algorithms and methods are needed to facilitate statistical inference at scale. Our particular focus has been on computational methods for Bayesian analyses, however we have encountered a number of general principles that will impact statistical computing research on Big Data problems. We have investigated how distributed computing, subsampling and randomised data compression can be used for efficient statistical computation.

A subtle yet fundamental distinction underpins how algorithms should be compared and assessed. This is the distinction between the number of floating point operations required by an algorithm and the typical execution time as measured by the end user (wall-clock time). Making such a differentiation is critical when allowing for parallel computation. A hardline accounting perspective is that the computational expense of an algorithm is defined by the number of floating point operations that are invoked, whether this is in parallel or not. A purely utilitarian viewpoint is that the only relevant factor is the time taken for the algorithm to return the end result, making wall clock-time the sole metric of interest.

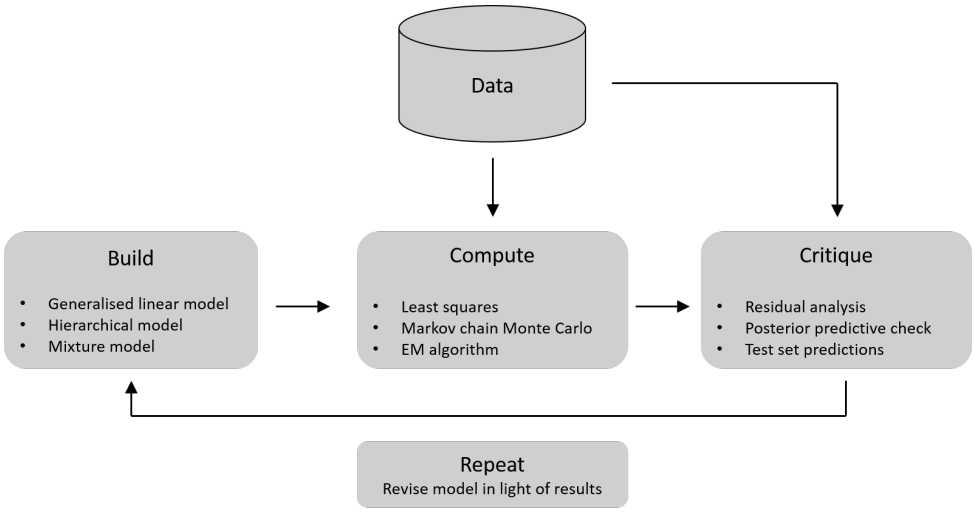


Figure 7.1: Box's loop is conceptual process model of scientific data analysis (Box, 1976). Box's loop defines a number of key phases (Build, compute, critique and repeat) when approaching a data modelling problem. The compartmentalisation of the overall task highlights important tactical decisions that are involved the statistical analysis lifecycle and aids the elicitation of broader strategic elements that influence the work. Adapted from Blei (2014).

Consider the task of calculating the mean of each variable given a large dataset of n observations on p variables. This can be executed on a single machine in $O(np)$ operations. It is trivial to define an embarrassingly algorithm for carrying out this task. Given that we distribute the job across s processors we can expect the wall-clock time to drop by a factor s assuming minimal communication costs. The wall-clock problem is solved by an embarrassingly parallel algorithm, however the raw number of floating point operations remains the same. Each of the s workers completes an $O(np/s)$ task leading to an overall expense of $O(np)$ operations. A pragmatic solution exists, yet the deeper issue of the floating point operation cost is not addressed by using a divide and conquer approach. The underlying expense of an algorithm is of interest in theoretical computer science. Reflective of this is that early work on sketching considered the estimation of a mean (Cormode, 2011). A more thorough analysis would of course include memory, disk and communication costs, we are choosing to emphasise the floating point operation count as it is of the most relevance to our work. If we can fundamentally reduce the number of floating point operations required in the compute step and still obtain an acceptable result, we have made great headway for numerical computing with Big Data. To surmise, subsampling and sketching cut wall clock time by reduce the underlying number of floating point operations. Divide and conquer approaches attack the wall clock time issue by distributing the floating point calculation burden over a large number of workers. A very classical statistical perspective may gloss over such implementation details, but this is an important aspect of dealing with Big Data (Donoho, 2017; Bryan and Wickham, 2017). The distinction between computational cost and wall-clock time should be kept in mind when comparing the work across chapters.

Specialised algorithms can be necessary for the practical analysis of large datasets. Parallel processing, subsampling and random projection can act as useful components within algorithms for Big data regression. Each broad approach has unique strengths and weaknesses and the best tool for the job may change on a case by case basis. To be more formal, recall the generic problem set-up from the introduction. Suppose we have a dataset $\boldsymbol{y}$ of n observations with likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$. The parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ has prior $p(\boldsymbol{\theta})$. Bayesian analyses can be challenging on tall datasets as computational procedures can be intolerably slow. The typical bottleneck encountered is the need for repeated full likelihood evaluations $p(\boldsymbol{y}|\boldsymbol{\theta})$, which carry an $O(n)$ cost. This can grow to an infeasible burden for a sufficiently large number of observations n (Robert et al., 2018). A main goal in the thesis to develop algorithms that minimise or eliminate $O(n)$ calculations. The divide and conquer work in Chapter 2 splits the initial dataset into s subsets so that each worker only has tasks that are $O(n/s)$. In the subsampling chapter we use subsamples of size $m \ll n$ to estimate the log likelihood in $O(m)$ time. In the work in sketching we replace the full dataset of n observations with a compressed pseudo-dataset of $k \ll n$ observations. Numerical work on the sketched dataset demands $O(k)$ operations.

Existing work in the field is largely directed at sampling from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$. To carry out model selection it is desirable to have the integrated likelihood (model evidence) $p(\boldsymbol{y}|\boldsymbol{\theta}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\ d\boldsymbol{\theta}$, and this was a motivating concern throughout the thesis. Chapter 2 uses divide and conquer methodology in order to compute the model evidence. Chapter3 uses subsampling to bound the model evidence. Chapters 4,5 and 6 examine how random projection can be used to approximate the least squares estimate $\widehat{\boldsymbol{\theta}}$ and the integrated likelihood $p(\boldsymbol{y})$ for Gaussian linear models.

The analysis of each technique required different mathematical methods and concepts. In Chapter 2 we developed an embarassingly parallel algorithm using Gibbs sampling and conditional conjugacy. The approach fits the split-apply-combine template shown in Figure 7.2. We propose to use Gibbs sampling in the apply stage, and to take the Gibbs sampler history as part of the output. The combine step involves pooling the Gibbs trajectories from each subset run. The initial partition of the data in the split step has a strong influence on the Monte Carlo variance in the combine step.

In Chapter 3 we motivated the use of subsampling for estimation of the integrated likelihood using

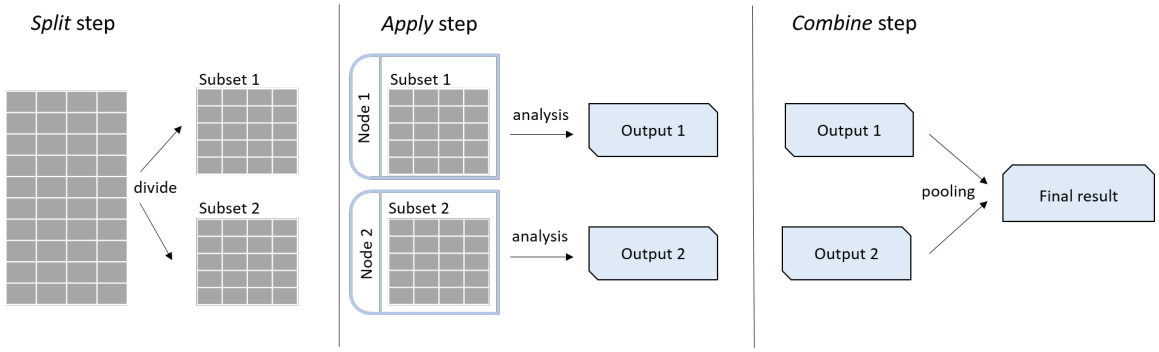


Figure 7.2: Template for embarrassingly parallel algorithms. The split step breaks the full dataset in to s non overlapping subsets. The illustration is for $s = 2$. Each subset is then allocated to a different machine on a cluster. During the *apply* step we apply conventional methodology to each data subset with no cross communication between workers. Each analysis is summarised by a consistent format of output. The s sets of output from the apply stage are then synthesised in the combine step to give the final result. In this design brief, the combine stage only involves the output from the apply step, and not the original dataset.

the the identity

$$\log p(\boldsymbol{y}) = \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{\theta})] - D(p(\boldsymbol{\theta}|\boldsymbol{y}) \parallel p(\boldsymbol{\theta})). \tag{7.1}$$

The goodness of fit term $\mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{\theta})]$ can be estimated using subsampling, and the penalty term $D(p(\boldsymbol{\theta}|\boldsymbol{y})\|p(\boldsymbol{\theta}))$ can be bounded using information theoretic arguments. The main contribution is an upper bound on the model evidence using the maximum entropy property of the normal distribution. Exact calculation of the model evidence becomes increasingly computationally demanding as n increases. Evidence bounds are attractive for large n problems as they can be expected to squeeze together as n increases under mild assumptions, and they can be estimated cheaply using control variate estimators.

Chapters 4, 5 and 6 considered sketching, a probabilistic data compression technique. As mentioned in the introduction, most of our work on sketching can be seen as a statistical evaluation of algorithms developed in the computer science and machine learning literature. We pried into the nature of the existing stochastic machinery behind randomised algorithms. Randomised algorithms are interesting from a statistical perspective as repeated application of the algorithm to the same dataset will produce different results. The distribution of the output is a measure of the quality of the algorithm. We established a number of asymptotic results regarding the distribution of the sketched output. In Chapter 4 we considered the distribution of least squares coefficients on the sketched dataset. In Chapter 6 we considered the probability of obtaining an ϵ-subspace embedding (Definition 6.1). Existing results on sketching algorithms are typically finite sample worst case bounds that can be pessimistic. We took a different approach and tried to characterise typical behaviour under regularity conditions. It is important to be able to quantify the uncertainty attached to a randomised algorithm, and the results in Chapter 4 on the variance of the sketched coefficients and the results in Chapter 6 about the embedding probability are useful measures for end users. We found that the central limit theorem and the Tracy-Widom law were useful for analysing the behaviour of data oblivious sketching algorithms. Weak convergence is a useful concept in the analysis of randomised algorithms. We believe our results serve as a useful complement to the existing body of work in the area. Sketching algorithms can be more richly analysed using a combination of non-asymptotic and asymptotic results.

The fundamental trade off between the computational budget and accuracy of the calculation is a primary concern for computational statistics in the Big Data world. This is in some sense no a free lunch problem, where lowering the computational expense of a Monte Carlo method can be expected to increase the Monte Carlo error. Successful Big Data algorithms use available resources wisely. Embarrassingly parallel algorithms spread the floating point operation cost over a large number of nodes without grossly increasing communication costs. Subsampling based algorithms use control variates as a stabilising influ-

ence. Sketching algorithms use well designed random projections to give robust performance guarantees in terms of the compression ratio and the structure of the source dataset. Truly scalable methodology controls the error in conjunction with the cost as datasets increase in size, and this is an overarching concern in each piece of work. Much of the statistical analysis in this thesis has been directed at non sampling based sources of error.

The Divide and conquer approach for computing the integrated likelihood (Chapter 2) provides an interesting example of this challenge. We easily achieve a linear reduction in the analysis time given that we split the full dataset into s subsets and process them in parallel. However the difficulty of the combine step increases with the number of subsets s. The variance of the kernel density estimator of the subposterior integral will be extremely high when s is large. As such, there is a limit to how much can be gained from parallel processing. Data augmentation can be used to address the scaling problem: by using Gibbs sampling in the intermediate apply step of the algorithm we can obtain a feasible combine step for large s. In Chapter 3, we find that although unbiased estimators of the log likelihood can be constructed using simple random sampling, the variance of such estimators explodes as n tends to infinity. Control variates are needed to obtain a scalable algorithm as n increases. A similar principle applies to sketching. A simple random sample of size k can be viewed as a sketch, however it does not offer the same probabilistic guarantees as the Hadamard and as Clarkson-Woodruff sketch n tends to infinity. Algorithms need to be designed carefully in order to control the associated Monte Carlo or approximation error.

There are a number of research avenues that appear to be worth pursing. As mentioned in Chapter 2, the split-apply combine approach using data augmentation and Gibbs sampling can also be used for posterior sampling. In Chapter 2 we considered estimation of the model evidence. We prescribe Gibbs sampling during the *apply* step and take the output to be the sequence of full conditional distributions. The *combine* step used the history of the subset Gibbs runs to estimate the integrated likelihood $p(\boldsymbol{y})$. We can also make use of the Gibbs history to target the full dataset posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ in the *combine* step. We can define a self-normalised importance sampler to target the true posterior distribution in the *combine* step using the providing that the model satisfies conditional conjugacy constraints.

The work on subsampling based estimation of the log model evidence in Chapter 3 can also be taken in new directions. Exact calculation of the integrated likelihood appears to require a large number of floating point operations. If this cost is inescapable, a useful strategy is to instead bound the integrated likelihood. We built our strategy around the identity (7.1). Subsampling can be used to estimated the goodness of fit term $\mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{\theta})]$. It is of interest to determine other cheap estimators of the Kullback-Leibler penalty term $D(p(\boldsymbol{\theta}|\boldsymbol{y})\|p(\boldsymbol{\theta}))$.

There was an important common pattern to the analysis in Chapters 4 and 6. Recall that the Gaussian sketch is particularly mathematically tractable but is computationally demanding whereas the Hadamard and Clarkson-Woodruff projections are computationally efficient but difficult to characterise mathematically. Our broad strategy was to analyse the Gaussian sketch and then establish asymptotic equivalence for the other projections. This was useful to develop guidelines for sketching applications. This general idea was suggested by Li et al. (2006), however it has appeared to have gained little traction in the sketching literature. We found this line of reasoning to be highly effective, and anticipate that it can be used address other questions concerning the behaviour of sketching algorithms. As in earlier chapters suppose we have a large $n \times d$ source dataset and we take a sketch of size k where $k \ll n$. It is of interest to extend the sketching central limit theorem to the 'Big Data' asymptotic regime:

$$n, d, k \to \infty, \quad d/k \to \alpha \in (0, 1], \quad d/n \to 0, \quad k/n \to 0. \tag{7.2}$$

In the regime (7.2) the sketched dataset becomes an infinite dimensional random matrix. It is possible to establish weak convergence in general metric spaces (Van Der Vaart, 1998), and we aim to identify the appropriate framework to study sketching under the limit conditions in (7.2). The concept of joint asymptotic normality as espoused by Mallows (1972) appears to be useful for establishing the necessary

convergence of the finite dimensional distributions. We anticipate that it will be necessary to treat the sketched dataset as a random bounded linear operator.

The connections between the Tracy-Widom law and the behaviour of data oblivious sketches may have deeper ramifications for statistical computation. Iterative sketching algorithms for least squares problems use a combination of approximate second order information from a sketches and exact gradient information. Iterative sketching algorithms can be interpreted as classical iterative methods with a random preconditioner (Gower and Richtrik, 2015; Pilanci and Wainwright, 2016). The iterates come with stochastic convergence guarantees to the optimal solution. It can be shown that the rate of convergence of the algorithm is governed by the eigenvalue distribution of the random preconditioner (Young, 1972). Knowledge of the spectrum can be used to accelerate the rate of convergence of the algorithm through the addition of learning rate and momentum hyperparameters (Young, 1972). We can implement Chebyshev acceleration on the sketched linear system given that we know the asymptotic spectrum of the random preconditioner. We are currently investigating the acceleration of iterative sketching algorithms using the spectral theory developed in Chapter 6.

The Bayesian analysis of tall data can be challenging due to obstacles encountered in the compute step of Box's loop. We believe distributed computing, subsampling and random projection are all promising approaches for computationally efficient Bayesian inference. We hope that the algorithms and results here expand the suite of computational tools for analysts to use in practice.

References

Ailon, N. and Chazelle, B. (2009) The fast Johnson Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, **39**, 302–322.

Anderson, I. (1997) *Combinatorial Designs and Tournaments*. Oxford lecture series in mathematics and its applications. Clarendon Press.

Andrieu, C., Roberts, G. O. et al. (2009) The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, **37**, 697–725.

Arnold, B. C., Castillo, E. and Sarabia, J. M. (1993) Conjugate exponential family priors for exponential family likelihoods. *Statistics*, **25**, 71–77.

Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M. A. et al. (2016) The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*, **167**, 1415–1429.

Avron, H., Nguyen, H. and Woodruff, D. (2014) Subspace embeddings for the polynomial kernel. In *Advances in Neural Information Processing Systems*, 2258–2266.

Bai, Z., Fang, Z., Liang, Y.-C. and Fange, Z. (2014) *Spectral theory of large dimensional random matrices*. Singapore ; Hackensack, N.J.

Bai, Z. and Silverstein, J. W. (2010) *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. New York, NY: Springer New York, 2nd edn.

Baker, J., Fearnhead, P., Fox, E. B. and Nemeth, C. (2017) Control variates for stochastic gradient MCMC. *arXiv preprint arXiv:1706.05439*.

Banerjee, A., Dunson, D. B. and Tokdar, S. T. (2013) Efficient Gaussian process regression for large datasets. *Biometrika*, **100**, 75–89.

Bao, Z., Pan, G. and Zhou, W. (2015) Universality for the largest eigenvalue of sample covariance matrices with general population. *The Annals of Statistics*, **43**, 382–421.

Bardenet, R., Doucet, A. and Holmes, C. (2014) Towards scaling up markov chain monte carlo: an adaptive subsampling approach. In *International Conference on Machine Learning (ICML)*, 405–413.

Bardenet, R., Doucet, A. and Holmes, C. (2015) On Markov chain Monte Carlo methods for tall data. *arXiv preprint 1505.02827*.

Bardenet, R., Doucet, A. and Holmes, C. (2017) On Markov chain Monte Carlo methods for tall data. *The Journal of Machine Learning Research*, **18**, 1515–1557.

Bardenet, R. and Maillard, O.-A. (2015) A note on replacing uniform subsampling by random projections in MCMC for linear regression of tall datasets. *HAL preprint 01248841*.

Becker, S., Kawas, B., Petrik, M. and Ramamurthy, K. (2015) Robust partially-compressed least- squares. *arXiv preprint arXiv:1510.04905v1*.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 289–300.

Bennett, C. H. (1976) Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, **22**, 245–268.

Bernardo, J., Bayarri, M. and Berger, J. (2011) *Bayesian Statistics 9*. Oxford science publications. OUP Oxford.

Bernardo, J. and Smith, A. (2006) *Bayesian Theory*. Wiley Series in Probability and Statistics. John Wiley & Sons Canada, Limited.

Besag, J. (1989) A candidate's formula: A curious result in bayesian prediction. *Biometrika*, **76**, 183.

Bhatia, R. (1996) *Matrix Analysis*. Springer.

Bierkens, J., Fearnhead, P. and Roberts, G. (2016) The zig-zag process and super-efficient sampling for Bayesian analysis of big data. *arXiv preprint arXiv:1607.03188*.

Billinglsley, P. (1968) *Convergence of probability measures*. Wiley.

Billingsley, P. (1999) *Convergence of probability measures*. Wiley Series in Probability and Statistics. New York: Wiley, 2nd edn.

Blei, D. M. (2014) Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, **1**, 203–232.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017) Variational inference: A review for statisticians. *Journal of the American Statistical Association*, **112**, 859–877.

Bottolo, L. and Richardson, S. (2010) Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis*, **5**, 583–618.

Box, G. E. P. (1976) Science and statistics. *Journal of the American Statistical Association*, **71**, 791–799.

Bryan, J. and Wickham, H. (2017) Data science: A three ring circus or a big tent? *Journal of Computational and Graphical Statistics*, **26**, 784–785.

Cannings, T. I. and Samworth, R. J. (2015) Random projection ensemble classification. *arXiv preprint arXiv:1504.04595*.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2017) Stan: A probabilistic programming language. *Journal of Statistical Software*, **76**.

Casella, G., Girón, F. J., Martínez, M. L., Moreno, E. et al. (2009) Consistency of Bayesian procedures for variable selection. *The Annals of Statistics*, **37**, 1207–1228.

Centers for Disease Control and Prevention (2013) Behavioral Risk Factor Surveillance System. Available online at `https://www.cdc.gov/brfss/annual_data/annual_2013.html`.

Chatterjee, D., Maitra, T. and Bhattacharya, S. (2018) A short note on almost sure convergence of bayes factors in the general set-up. *The American Statistician*, **0**, 1–4.

Chib, S. (1995) Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, **90**, 1313–1321.

Chib, S. and Kuffner, T. A. (2016) Bayes factor consistency. *arXiv preprints arXiv:1607.00292*.

Clarkson, K. L. and Woodruff, D. P. (2013) Low rank approximation and regression in input sparsity time. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, 81–90. ACM.

Cormode, G. (2011) Sketch techniques for approximate query processing. *Foundations and Trends in Databases*.

Dempster, A. P. (1997) The direct use of likelihood for significance testing. *Statistics and Computing*, **7**, 247–252.

Dhillon, P., Lu, Y., Foster, D. P. and Ungar, L. (2013) New subsampling algorithms for fast least squares regression. In *Advances in Neural Information Processing Systems*, 360–368.

DiCiccio, T. J., Kass, R. E., Raftery, A. and Wasserman, L. (1997) Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, **92**, 903–915.

Dickey, J. M. (1971) The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, 204–223.

Donoho, D. (2017) 50 years of data science. *Journal of Computational and Graphical Statistics*, **26**, 745–766.

Drineas, P., Mahoney, M. W. and Muthukrishnan, S. (2006) Sampling algorithms for l2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, 1127–1136. Society for Industrial and Applied Mathematics.

Eaton, M. L. (2007) Chapter 8: The Wishart Distribution. In *Multivariate Statistics*, vol. 53 of *Lecture Notes Monograph Series*, 302–333. Ohio: Institute of Mathematical Statistics.

Edelman, A. (1988) Eigenvalues and condition numbers of random matrices. *SIAM Journal on Matrix Analysis and Applications*, **9**, 543–560.

Efron, B. and Hastie, T. (2016) *Computer Age Statistical Inference*. Institute of Mathematical Statistics Monographs. Cambridge University Press.

Fahrmeir, L. and Tutz, G. (1994) *Multivariate statistical modelling based on generalized linear models*. Springer series in statistics. Springer-Verlag.

Flegal, J. M., Hughes, J., Vats, D. and Dai, N. (2017) *mcmcse: Monte Carlo Standard Errors for MCMC*. Riverside, CA, Denver, CO, Coventry, UK, and Minneapolis, MN. R package version 1.3-2.

Friel, N. and Pettitt, A. N. (2008) Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 589–607.

Friel, N. and Wyse, J. (2012) Estimating the evidence a review. *Statistica Neerlandica*, **66**, 288–308.

Gelfand, A. E. and Dey, D. K. (1994) Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 501–514.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2014) *Bayesian data analysis*. Boca Raton: Chapman & Hall, 3 edn.

Gelman, A. and Vehtari, A. (2017) Comment: Consensus Monte Carlo using expectation propagation. *Brazillian Journal of Probability and Statistics*, **31**, 692–696.

Geman, S. (1980) A limit theorem for the norm of random matrices. *The Annals of Probability*, 252–261.

Geppert, L. N., Ickstadt, K., Munteanu, A., Quedenfeld, J. and Sohler, C. (2017) Random projections for Bayesian regression. *Statistics and Computing*, **27**, 79–101.

Geyer, C. (1996) Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. *Tech. Rep. 568*, University of Minnesota.

Gower, R. M. and Richtrik, P. (2015) Randomized iterative methods for linear systems. *arXiv preprint arXiv:1506.03296*.

Green, P. J., Łatuszyński, K., Pereyra, M. and Robert, C. P. (2015) Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, **25**, 835–862.

Greene, W. (1997) *Econometric Analysis*. Prentice-Hall international editions. Prentice Hall.

Guhaniyogi, R. and Dunson, D. B. (2015) Bayesian compressed regression. *Journal of the American Statistical Association*, **110**, 1500–1514.

Gutiérrez-Peña, E., Smith, A., Bernardo, J. M., Consonni, G., Veronese, P., George, E., Girón, F., Martínez, M., Letac, G. and Morris, C. N. (1997) Exponential and Bayesian conjugate families: review and extensions. *Test*, **6**, 1–90.

Halko, N., Martinsson, P. G. and Tropp, J. A. (2011) Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, **53**, 217–288.

Howie, B. N., Donnelly, P. and Marchini, J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, **5**, e1000529.

Huang, Z. and Gelman, A. (2005) Sampling for Bayesian computation with large datasets. *Tech. rep.*, University of Columbia.

Huber, P. J. (1973) Robust regression: Asymptotics, conjectures and monte carlo. *The Annals of Statistics*, **1**, 799–821. URL`https://doi.org/10.1214/aos/1176342503`.

Jacob, P. E., Thiery, A. H. et al. (2015) On nonnegative unbiased estimators. *The Annals of Statistics*, **43**, 769–784.

Jahn, R., Dunne, B. and Nelson, R. (1987) Engineering anomalies research. *Journal of Scientific Exploration*.

Janson, S. (1988) Some pairwise independent sequences for which the central limit theorem fails. *Stochastics*, **23**, 439–448.

Johnstone, I. M. (2001) On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, 295–327.

— (2006) High dimensional statistical inference and random matrices. *arXiv preprint arXiv:0611589*.

Johnstone, I. M., Ma, Z., Perry, P. O. and Shahram, M. (2014) *RMTstat: Distributions, Statistics and Tests derived from Random Matrix Theory*. R package version 0.3.

Jordan, M. I. (2013) On statistics, computation and scalability. *Bernoulli*, **19**, 1378–1390.

Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Keener, R. W. (2013) *Theoretical Statistics*. Springer.

Korattikara, A., Chen, Y. and Welling, M. (2014) Austerity in mcmc land: Cutting the metropolis-hastings budget. In *International Conference on Machine Learning*, 181–189.

Kuhn, M. (2008) Building predictive models in R using the caret package. *Journal of Statistical Software, Articles*, **28**, 1–26.

Lehoucq, R. B., Sorensen, D. C. and Yang, C. (1998) *ARPACK users' guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*, vol. 6. Siam.

Lewis, S. M. and Raftery, A. E. (1997) Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *Journal of the American Statistical Association*, **92**, 648–655.

Li, P., Hastie, T. J. and Church, K. W. (2006) Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 287–296. ACM.

Loeve, M. (1977) *Probability Theory*. Springer.

Ma, P., Mahoney, M. W. and Yu, B. (2015) A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, 861–911.

Ma, P. and Sun, X. (2015) Leveraging for big data regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, **7**, 70–76.

Ma, Z. (2012) Accuracy of the Tracy–Widom limits for the extreme eigenvalues in white Wishart matrices. *Bernoulli*, **18**, 322–359.

Maclaurin, D. and Adams, R. P. (2014) Firefly Monte Carlo: Exact MCMC with Subsets of Data. *arXiv preprint arXiv:1403.5693*.

Mahoney, M. (2011) Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, **3**, 123–224.

Mahoney, M. and Drineas, P. (2016) Structural properties underlying high-quality Randomized Numerical Linear Algebra algorithms. In *Handbook of Big Data* (eds. P. Buhlmann, P. Drineas, M. Kane and M. van de Laan), 137–154. Chapman and Hall.

Mallows, C. (1972) A note on asymptotic joint normality. *The Annals of Mathematical Statistics*, 508–515.

McCulloch, R. E. and Rossi, P. E. (1992) Bayes factors for nonlinear hypotheses and likelihood distributions. *Biometrika*, **79**, 663–676.

Meng, X. (2014) *Randomized Algorithms for Large-scale Strongly Over-determined Linear Regression Problems*. Ph.D. thesis, Stanford University, Stanford, California, United States.

Meng, X. and Mahoney, M. M. (2013) Low-distortion Subspace Embeddings in Input-sparsity Time and Applications to Robust Linear Regression. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, 91–100. ACM.

Meng, X.-L. and Wong, W. H. (1996) Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 831–860.

Minsker, S., Srivastava, S., Lin, L. and Dunson, D. B. (2017) Robust and scalable Bayes via a median of subset posterior measures. *The Journal of Machine Learning Research*, **18**, 4488–4527.

Neiswanger, W., Wang, C. and Xing, E. (2013) Asymptotically exact, embarrassingly parallel MCMC. *arXiv preprint arXiv:1311.4780*.

Nelson, J. and Nguyên, H. L. (2013) Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, 117–126. IEEE.

Newton, M. A. and Raftery, A. E. (1994) Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, 3–48.

Papaspiliopoulos, O. (2009) A methodological framework for Monte Carlo probabilistic inference for diffusion processes. *Working paper*, University of Warwick, Coventry.

Phillips, J. M. (2016) Coresets and Sketches. *arXiv preprint arXiv:1601.00617*.

Pilanci, M. and Wainwright, M. J. (2016) Iterative Hessian sketch: Fast and accurate solution approximation for constrained least-squares. *Journal of Machine Learning Research*, **17**, 1842–1879.

Pollock, M., Fearnhead, P., Johansen, A. M. and Roberts, G. O. (2016) The scalable Langevin exact algorithm: Bayesian inference for big data. *arXiv preprint arXiv:1609.03436*.

Polson, N. G., Scott, J. G. and Windle, J. (2013) Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, **108**, 1339–1349.

Portnoy, S. (1986) On the central limit theorem in rp when p tends to infinity. *Probability Theory and Related Fields*, **73**, 571–583.

Pruss, A. R. and Szynal, D. (2000) On the central limit theorem for negatively correlated random variables with negatively correlated squares. *Stochastic Processes and their Applications*, **87**, 299 – 309.

Quiroz, M., Kohn, R., Villani, M. and Tran, M.-N. (2018) Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association*, **0**, 1–35.

Quiroz, M., Tran, M.-N., Villani, M., Kohn, R. and Dang, K.-D. (2016) The block-Poisson estimator for optimally tuned exact subsampling MCMC. *ArXiv e-prints*.

Raftery, A. E. (1995) Bayesian model selection in social research. *Sociological Methodology*, **25**, 111–163.

— (1996) Hypothesis testing and model selection via posterior simulation. In *Markov chain Monte Carlo in practice*, 163–188.

Raskutti, G. and Mahoney, M. (2014) A Statistical Perspective on Randomized Sketching for Ordinary Least-Squares. *arXiv preprint arXiv:1406.5986*.

Rhee, C.-h. and Glynn, P. W. (2015) Unbiased estimation with square root convergence for SDE models. *Operations Research*, **63**, 1026–1043.

Robert, C. P. (2007) *The Bayesian Choice From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. New York, NY: Springer New York, 2nd edn.

Robert, C. P. and Casella, G. (2010) *Monte Carlo Statistical Methods*. Springer.

Robert, C. P., Elvira, V., Tawn, N. and Wu, C. (2018) Accelerating MCMC algorithms. *Wiley Interdisciplinary Reviews: Computational Statistics*, **10**, 1–14.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. and Mller, M. (2011) proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, **12**, 77.

Roosta-Khorasani, F. and Mahoney, M. W. (2016) Sub-Sampled Newton Methods I: Globally Convergent Algorithms. *arXiv preprint arXiv:1601.04737*.

Sarlos, T. (2006) Improved approximation algorithms for large matrices via random projections. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, 143–152. IEEE.

Sarndal, C.-E., Swensson, B. and Wretman, J. H. (1992) *Model assisted survey sampling*. Springer series in statistics. New York: Springer-Verlag.

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Scott, S. L. (2017) Comparing consensus Monte Carlo strategies for distributed Bayesian computation. *Brazilian Journal of Probability and Statistics*, **31**, 668–685.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H., George, E. and McCulloch, R. (2013) Bayes and big data: the consensus Monte Carlo algorithm. In *EFaBBayes 250 conference*, vol. 16.

Searle, S. R. (1997) *Linear Models*. New Jersey: Wiley-Interscience.

Shah, R. D. and Meinshausen, N. (2013) Min-wise hashing for large-scale regression and classification with sparse data. *arXiv preprint arXiv:1308.1269*.

Shirts, M. R., Bair, E., Hooker, G. and Pande, V. S. (2003) Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods. *Physical Review Letters*, **91**, 140601.

Silverman, B. W. (1986) *Density estimation for statistics and data analysis*. Monographs on statistics and applied probability (Series). London ; New York: Chapman and Hall.

Silverstein, J. W. (1985) The smallest eigenvalue of a large dimensional wishart matrix. *The Annals of Probability*, **13**, 1364–1368. URL`https://doi.org/10.1214/aop/1176992819`.

Skilling, J. et al. (2006) Nested sampling for general Bayesian computation. *Bayesian Analysis*, **1**, 833–859.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 583–639.

Srivastava, S., Cevher, V., Tran-Dinh, Q., Dunson, D. B. and de Lausanne, F. (2015) WASP: Scalable Bayes via barycenters of subset posteriors. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 912–920.

Stan Development Team (2018) *Stan Modeling Language Users Guide and Reference Manual*.

Tanner, M. A. and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–540.

Terrell, G. R. and Scott, D. W. (1992) Variable kernel density estimation. *The Annals of Statistics*, **20**, 1236–1265.

Thanei, G.-A., Heinze, C. and Meinshausen, N. (2017) Random projections for large-scale regression. *arXiv preprint 1701.05325*.

Tracy, C. A. and Widom, H. (1994) Level-spacing distributions and the airy kernel. *Communications in Mathematical Physics*, **159**, 151–174.

Tropp, J. A. (2011) Improved analysis of the subsampled randomized Hadamard transform. *Advances in Adaptive Data Analysis*, **3**, 115–126.

Van Der Vaart, A. (1998) *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, 3. Cambridge University Press.

Varian, H. R. (2014) Big data: New tricks for econometrics. *Journal of Economic Perspectives*, **28**, 3–28.

Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S*. New York: Springer, fourth edn. ISBN 0-387-95457-0.

Venkatasubramanian, S. and Wang, Q. (2011) The Johnson-Lindenstrauss transform: an empirical study. In *2011 Proceedings of the Thirteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, 164–173. SIAM.

Wagner, W. (1987) Unbiased Monte Carlo evaluation of certain functional integrals. *Journal of Computational Physics*, **71**, 21–33.

Walker, S., Damien, P. and Lenk, P. (2004) On priors with a Kullback-Leibler property. *Journal of the American Statistical Association*, **99**, 404–408.

Walschap, G. (2015) *Multivariable Calculus and Differential Geometry*. De Gruyter.

Wang, X. and Dunson, D. B. (2013) Parallelizing MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*.

Wasserstein, R. L. and Lazar, N. A. (2016) The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, **70**, 129–133.

Welling, M. and Teh, Y. W. (2011) Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 681–688.

Wetzels, R., Grasman, R. P. and Wagenmakers, E.-J. (2010) An encompassing prior generalization of the Savage–Dickey density ratio. *Computational Statistics & Data Analysis*, **54**, 2094–2102.

White, H. (1984) *Asymptotic theory for econometricians*. Economic Theory, Econometrics and Mathematical Economics Series. Academic Press.

Wickham, H. (2014) *nycflights13: Data about flights departing NYC in 2013*. R package version 0.1.

Woodruff, D. P. (2014) Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, **10**, 1–157.

Yang, J., Meng, X. and Mahoney, M. W. (2015a) Implementing randomized matrix algorithms in parallel and distributed environments. *arXiv preprint arXiv:1502.03032*.

Yang, T., Zhang, L., Lin, Q. and Jin, R. (2015b) Fast sparse least-squares regression with non-asymptotic guarantees. *arXiv preprint arXiv:1507.05185*.

Young, D. M. (1972) Second-degree iterative methods for the solution of large linear systems. *Journal of Approximation Theory*, **5**, 137 – 148.

Zhou, S., Ligett, K. and Wasserman, L. (2009) Differential privacy with compression. In *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, 2718–2722. IEEE.